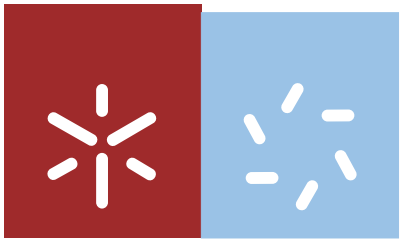
Universidade do Minho
Escola de Ciências

Artur Agostinho Marinho de Araújo

Estimação da função de distribuição bivariada na presença de censura

Outubro de 2012

Estimação da função de distribuição bivariada na presença de censura
Artur Agostinho Marinho de Araújo
UMinho | 2012



Universidade do Minho
Escola de Ciências

Artur Agostinho Marinho de Araújo

Estimação da função de distribuição bivariada na presença de censura

Dissertação de Mestrado
Mestrado em Estatística

Trabalho realizado sob a orientação do
Professor Doutor Luís Filipe Meira Machado

Outubro de 2012

Agradecimentos

Em primeiro lugar gostaria de agradecer ao meu pai e à minha mãe, pois sem eles eu não teria frequentado o Mestrado em Estatística na Universidade do Minho.

Gostaria de agradecer ao Doutor Luís Filipe Meira Machado pela sua orientação no trabalho de investigação que deu origem a esta dissertação. Pelas muitas indicações de bibliografia adequada. Por todo o trabalho que teve ao ler as muitas revisões desta dissertação, sugerindo a cada revisão a introdução de novos melhoramentos.

Uma palavra de agradecimento a todos aqueles que contribuíram para a ciência nos seus domínios relacionados com este trabalho, nomeadamente a Matemática, a Estatística e a Computação. Gostaria também de agradecer a todos aqueles que introduziram inovações tecnológicas, sobretudo nas áreas da Electrónica e da Computação. Gostaria ainda de agradecer a todos os que trabalham nas fábricas que produzem as máquinas que todos nós utilizamos no processamento de informação. Sem todas essas pessoas este trabalho não teria sido realizado.

Este trabalho foi financiado por fundos FEDER através do Programa Operacional Factores de Competitividade - COMPETE e por fundos nacionais através da FCT - Fundação para a Ciência e a Tecnologia no âmbito do projecto PTDC/MAT/104879/2008.

Artur Agostinho Marinho de Araújo

Esta página foi intencionalmente deixada em branco.

Estimação da função de distribuição bivariada na presença de censura

Resumo

Em muitos estudos longitudinais, os indivíduos experimentam eventos recorrentes. Este tipo de dados é frequentemente observado em investigação médica, em engenharia e áreas afins. Dados dois tempos provenientes de eventos sequenciais, vários métodos foram desenvolvidos para a estimação de quantidades de interesse, como a função de distribuição bivariada, as funções de distribuição marginais ou a função de distribuição condicional. Na maioria destes métodos admite-se que os indivíduos que entram no estudo continuam em observação até ao fim do estudo. No entanto em muitas aplicações isso não ocorre, e os indivíduos não são observados até ao final do estudo por várias razões. Neste trabalho é estudada a função de distribuição bivariada na presença de censura pela direita.

Relativamente à organização da presente dissertação, a mesma foi dividida em cinco secções.

Na primeira secção são resumidos os fundamentos teóricos mais relevantes de forma a facilitar ao leitor com alguns conhecimentos nos domínios da Matemática ou da Estatística a compreensão das secções seguintes. Assim são introduzidos conceitos fundamentais nas áreas da análise de sobrevivência, da estimação pontual e da estatística multivariada.

Na segunda secção é introduzido o problema da estimação da função de distribuição bivariada para tempos sequenciais na presença de censura pela direita. Quatro estimadores distintos são apresentados.

A terceira secção apresenta uma extensão para o *software* estatístico R, desenvolvida no âmbito da investigação que esteve na origem da escrita desta dissertação. A extensão mencionada foi desenvolvida com o objectivo de proporcionar resultados numéricos e gráficos em relação a cada um dos quatro estimadores apresentados nesta dissertação. De modo a demonstrar as potencialidades da extensão desenvolvida, são apresentados exemplos de utilização numérica e gráfica, acompanhados de possíveis interpretações.

A quarta secção descreve detalhadamente um estudo de simulação envolvendo os quatro estimadores abordados nesta dissertação. Nesta secção cada um dos estimadores é analisado em relação às propriedades desejáveis das funções de distribuição bivariadas. O comportamento de cada um dos estimadores em relação ao enviesamento, ao desvio padrão e ao erro quadrático médio é estudado recorrendo a gráficos e tabelas. De modo a identificar o estimador mais eficiente, os quatro estimadores são comparados dois a dois em termos da eficiência relativa.

Na quinta e última secção são apresentadas as principais conclusões obtidas do estudo de simulação descrito na secção anterior.

Estimation of the bivariate distribution function under censoring

Abstract

In many longitudinal studies, the individuals experience recurrent events. This type of data is frequently observed in medical research, engineering and related fields. Given two times coming from sequential events, several methods have been developed for the estimation of quantities of interest, such as the bivariate distribution function, the marginal distribution functions, and the conditional distribution function. Most of these methods admit that the individuals entering the study remain in observation until the end of the study. However, in many applications that doesn't happen, and the individuals aren't observed until the end of the study for many reasons. In this work the bivariate distribution function under right censoring is studied.

Pertaining to the organization of the present dissertation, the same was divided in five sections.

In the first section the most relevant theoric fundaments are resumed. So that a reader with some background in Mathematics or Statistics can easily understand the contents of the following sections. In this way fundamental concepts in the areas of survival analysis, point estimation and multivariate statistics are introduced.

The second section introduces the problem of estimation of the bivariate distribution function for sequentially ordered events under right censoring. Four distinct estimators are presented.

In the third section, a package for the statistical software R, developed in the framework of the research that led to this dissertation, is presented. This software package was developed with the purpose of producing numerical and graphical output for each of the four estimators discussed in this dissertation. To demonstrate some of the possibilities of the software package, numerical and graphical examples are given accompanied by possible interpretations.

The fourth section describes the details of a simulation study involving each one of the four estimators discussed in this dissertation. Each one of the estimators is analyzed in respect of the desirable properties of the bivariate distribution functions. With the aid of graphics and tables, the

behavior of each one of the estimators is studied in respect to bias, standard deviation and mean square error. To identify the most efficient estimator, the four estimators are compared by means of the relative efficiency.

The fifth and last section presents the main conclusions of the simulation study described in the previous section.

Índice

1 Introdução

Antes da abordagem do tema principal desta dissertação, serão introduzidos alguns conceitos fundamentais relevantes para uma adequada compreensão do mesmo tema.

É importante referir que neste texto, variáveis aleatórias são representadas por letras maiúsculas, e variáveis determinísticas são representadas por letras minúsculas, o que inclui realizações de variáveis aleatórias.

Ao longo deste capítulo, a variável aleatória T representa "tempo até ocorrer o evento". O evento pode ser qualquer acontecimento que possa ser definido. Em aplicações reais de análise de sobrevivência, os eventos mais comuns podem ser: falha de um equipamento, morte de um indivíduo, recorrência dos sintomas de uma doença, etc. A variável t representa uma realização da variável aleatória T.

1.1 Função de sobrevivência

A quantidade básica empregue na descrição de fenómenos "tempo até ocorrer o evento", é a função de sobrevivência, a probabilidade de um individuo sobreviver para além de um tempo t. É definida como:

$$S(t) = P[T > t] \tag{1.1.1}$$

Se T for uma variável aleatória contínua, então $S(t)$ é uma função contínua monótona decrescente. A função de sobrevivência é o complemento da função de distribuição cumulativa ou seja:

$$S(t) = 1 - F(t) \tag{1.1.2}$$

onde:

$$F(t) = P[T \leq t] \tag{1.1.3}$$

A função de sobrevivência relaciona-se com a função densidade de probabilidade $f(u)$ segundo a equação:

$$S(t) = P[T > t] = \int_t^{+\infty} f(u)du \tag{1.1.4}$$

Pelo que:

$$f(t) = -\frac{dS(t)}{dt} \tag{1.1.5}$$

Note-se que $f(t)$ é uma função não negativa com a área abaixo da curva definida por $f(t)$ igual à unidade.

Quanto T é uma variável aleatória discreta é apropriado definir a função de sobrevivência de outra forma. Suponha-se que T pode tomar valores $t_i, i = 1, 2, \ldots$ com função massa de probabilidade $p(t_i) = P[T = t_i], t = 1, 2, \ldots$, onde $t_1 < t_2 < \cdots$. A função de sobrevivência para uma variável aleatória discreta relaciona-se com a função massa de probabilidade:

$$S(t) = P[T > t] = \sum_{t_i > t} p(t_i) \tag{1.1.6}$$

Note-se que quando T é discreta, a função de sobrevivência é uma função escada não crescente. A função massa de probabilidade pode ser obtida a partir da função de sobrevivência recorrendo à equação:

$$p(t_i) = S(t_{i-1}) - S(t_i) \tag{1.1.7}$$

A função de distribuição cumulativa para uma variável aleatória discreta pode ser escrita:

$$F(t) = P[T \leq t] = \sum_{t_i \leq t} p(t_i) = 1 - \sum_{t_i > t} p(t_i) \tag{1.1.8}$$

onde para obter a última igualdade se recorreu às equações (1.1.2) e (1.1.6). Note-se ainda que a função de sobrevivência pode ser escrita como o produto de probabilidades de sobrevivência condicionais:

$$\begin{aligned} S(t_i) &= \frac{S(t_i)}{S(t_{i-1})}\frac{S(t_{i-1})}{S(t_{i-2})}\ldots\frac{S(t_2)}{S(t_1)}\frac{S(t_1)}{S(t_0)}S(t_0) \\ &= P[T > t_i | T \geq t_i]P[T > t_{i-1} | T \geq t_{i-1}] \ldots \\ &\quad P[T > t_2 | T \geq t_2]P[T > t_1 | T \geq t_1] \end{aligned} \tag{1.1.9}$$

onde $S(t_0) = 1$. Pelo que a mesma equação pode ser escrita de forma mais compacta:

$$S(t) = \prod_{t_i \leq t} \frac{S(t_i)}{S(t_{i-1})} \tag{1.1.10}$$

1.2 Função de risco

A função de risco não é relevante para a compreensão do tema desta dissertação. No entanto trata-se de uma função de extrema importância em análise de sobrevivência. Pode ser relacionada com a função de sobrevivência. Para variáveis aleatórias discretas a relação entre a função de sobrevivência e a função de risco tem uma semelhança evidente com o estimador de Kaplan-Meier para a função de sobrevivência, como adiante se verá. Este estimador não paramétrico é fundamental para a estimação não paramétrica da função de distribuição bivariada na presença de censura. Pelas razões mencionadas opta-se por fazer uma breve referência à função de risco.

A função de risco é definida pela equação:

$$h(t) = \lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t | T \geq t]}{\Delta t} \tag{1.2.1}$$

A partir da equação (1.2.1) pode-se verificar que a quantidade $h(t)\Delta t$ pode ser vista como a probabilidade "aproximada" de um individuo com idade t experienciar o evento no instante imediato. Se T for uma variável aleatória contínua então:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{dln[S(t)]}{dt} \tag{1.2.2}$$

Uma quantidade relacionada trata-se da função de risco cumulativa $H(t)$ assim definida:

$$H(t) = \int_0^t h(u)du = -\ln[S(t)] \tag{1.2.3}$$

Assim para variáveis aleatórias contínuas, a função de sobrevivência relaciona-se com as funções de risco cumulativa e função de risco de acordo com a equação:

$$S(t) = e^{-H(t)} = e^{-\int_0^t h(u)du} \tag{1.2.4}$$

A função de risco pode tomar muitas formas, sendo que a única restrição imposta é a não negatividade, i.e., $h(t) \geq 0$.

Quando T é uma variável aleatória discreta, a função de risco é dada pela equação:

$$h(t_i) = P[T = t_i | T \geq t_i] = \frac{p(t_i)}{S(t_{i-1})}, i = 1, 2, \ldots \quad (1.2.5)$$

onde $S(t_0) = 1$. A equação (1.2.5) em conjunto com a equação (1.1.7) resulta:

$$h(t_i) = 1 - \frac{S(t_i)}{S(t_{i-1})}, i = 1, 2, \ldots \quad (1.2.6)$$

Lembre-se a equação (1.1.10) onde a função de sobrevivência é escrita como o produto das probabilidades de sobrevivência condicionais. Então, a função de sobrevivência está relacionada com a função de risco pela equação:

$$S(t) = \prod_{t_i \leq t} [1 - h(t_i)] \quad (1.2.7)$$

Note-se que para variáveis aleatórias discretas, a função risco é igual a zero, excepto em pontos onde uma falha ocorre. Para tempos de vida discretos, a função de risco cumulativa define-se:

$$H(t) = \sum_{t_i \leq t} h(t_i) \quad (1.2.8)$$

Note-se que a relação (1.2.4) não se mantém verdadeira para esta definição. Pelo que alguns autores preferem definir a função de risco cumulativa para tempos de vida discretos, da seguinte forma:

$$H(t) = -\sum_{t_i \leq t} \ln[1 - h(t_i)] \quad (1.2.9)$$

Pois a relação (1.2.4) válida para variáveis aleatórias contínuas torna-se válida também para variáveis aleatórias discretas. Se os valores $h(t_i)$ forem pequenos a equação (1.2.8) será uma aproximação à equação(1.2.9). A equação (1.2.8) é preferida, porque pode ser directamente estimada a partir de amostras com tempos de vida censurados ou truncados, e o estimador apresenta propriedades estatísticas desejáveis.[1]

Uma representação gráfica das referidas funções para um modelo exponencial de parâmetro igual a 2 pode ser visualizada na Figura 1.1.

[1] Trata-se do estimador de Nelson-Aalen para a função de risco cumulativa. Para mais detalhes veja-se por exemplo (Klein & Moeschberger, 2003, p. 94) ou ainda (Hosmer, Jr. & Lemeshow, 1999, pp. 73-77).

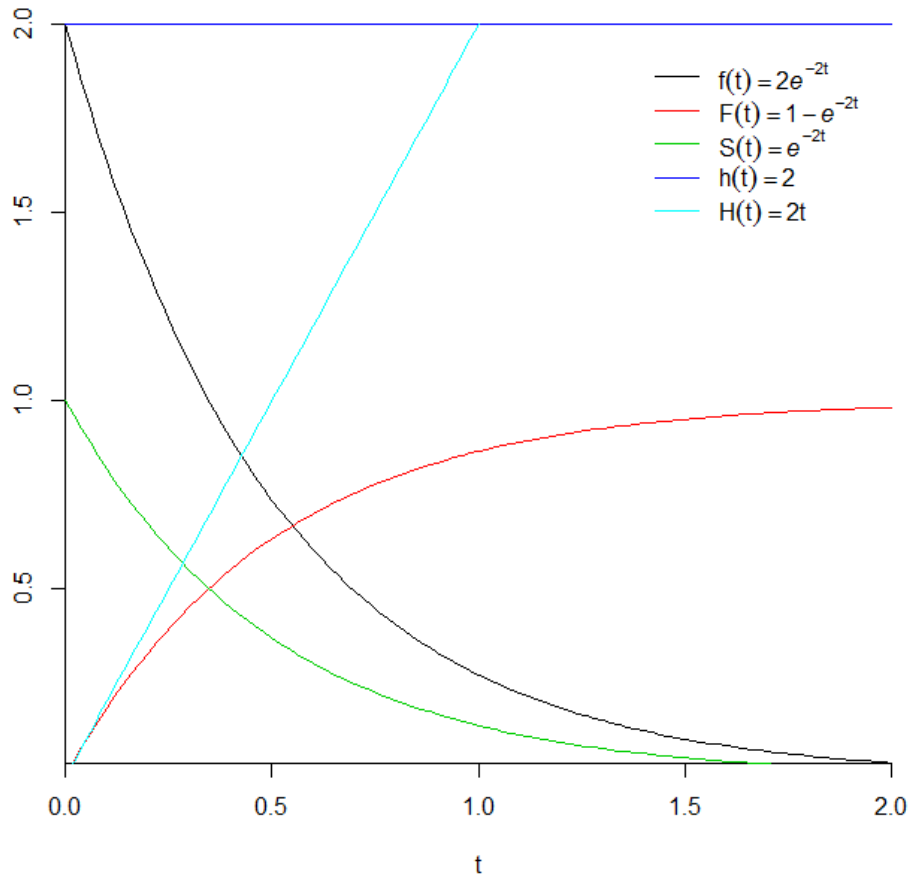


Figura 1.1: Diversas representações gráficas do modelo exponencial de parâmetro igual a 2.

O modelo exponencial é um modelo paramétrico contínuo, frequentemente empregue para modelar fenómenos que envolvem tempo até ocorrer um evento de interesse. Este modelo caracteriza-se por ter uma função de risco constante e igual ao parâmetro do próprio modelo. A função de risco cumulativa é uma recta de declive positivo e igual ao parâmetro do modelo. Estas representações gráficas relativamente simples das funções de risco e de risco cumulativa do modelo exponencial são usadas na prática para verificar se o modelo exponencial se ajusta a uma determinada amostra, bem como para estimar o parâmetro do modelo.

1.3 Estimação pontual

Será feita uma breve referência a alguns conceitos sobre teoria de estimação pontual. Tal teoria revela-se importante no contexto do presente trabalho, pois proporciona critérios que permitem comparar diversos estimadores de um mesmo parâmetro. Sendo um dos objectivos deste trabalho, a comparação de vários estimadores para a função de distribuição bivariada de tempos sequenciais na presença de censura. Para um estudo mais aprofundado da mesma teoria, sugere-se a leitura de (Knight, 2000) e ou (Lehmann & Casella, 1998).

Um estimador pontual ou estimador é uma estatística, cujo objectivo principal é a estimação de um parâmetro. Assuma-se que θ é um parâmetro e que $\hat{\theta}$ é um estimador de θ. Tratando-se o estimador $\hat{\theta}$ de uma estatística, será função de uma amostra; estando por isso sujeito a erro. A distribuição de probabilidade de um estimador $\hat{\theta}$ é referida por distribuição amostral de $\hat{\theta}$. Idealmente a distribuição amostral de $\hat{\theta}$ deve estar concentrada muito perto do verdadeiro valor do parâmetro θ. Foram definidas algumas medidas simples da qualidade de um estimador, baseadas na sua distribuição amostral. Tais medidas apresentam interesse, pois permitem em princípio descartar alguns estimadores, ou seleccionar uns estimadores em detrimento de outros.

O viés ou enviesamento de um estimador é definido por:

$$Env\left[\hat{\theta}\right] = E\left[\hat{\theta}\right] - \theta \tag{1.3.1}$$

onde $E\left[\hat{\theta}\right]$ representa a esperança de $\hat{\theta}$. Um estimador $\hat{\theta}$ é não enviesado ou centrado se $Env\left[\hat{\theta}\right] = 0$, ou seja $E\left[\hat{\theta}\right] = \theta$.

O erro quadrático médio de $\hat{\theta}$ define-se por:

$$EQM\left[\hat{\theta}\right] = E\left[(\hat{\theta} - \theta)^2\right] \tag{1.3.2}$$

Pode ser facilmente demonstrado que:

$$EQM\left[\hat{\theta}\right] = Var\left[\hat{\theta}\right] + \left[Env\left[\hat{\theta}\right]\right]^2 \tag{1.3.3}$$

onde $Var[\hat{\theta}]$ representa a variância do estimador $\hat{\theta}$. Suponha-se que $\theta_1, \dots, \theta_n$ são n realizações da variável aleatória $\hat{\theta}$, então $EQM\left[\hat{\theta}\right]$ pode ser estimado por:

$$EQM_{\hat{\theta}} = \frac{1}{n}\sum_{i=1}^{n}(\theta_i - \theta)^2 \tag{1.3.4}$$

A esperança de $\hat{\theta}$ pode ser facilmente estimada recorrendo à média amostral:

$$\bar{X}_{\hat{\theta}} = \frac{1}{n}\sum_{i=1}^{n}\theta_i \tag{1.3.5}$$

A variância de $\hat{\theta}$ por sua vez, pode ser estimada recorrendo à variância amostral:

$$S^2{}_{\hat{\theta}} = \frac{1}{n-1}\sum_{i=1}^{n}(\theta_i - \bar{X}_{\hat{\theta}})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}\theta_i{}^2 - n\bar{X}_{\hat{\theta}}{}^2\right] \tag{1.3.6}$$

Então, pelas equações (1.3.3) e (1.3.1), o estimador do erro quadrático médio descrito pela equação (1.3.4), pode ser reescrito:

$$EQM_{\hat{\theta}} = \frac{n-1}{n}S^2{}_{\hat{\theta}} + (\bar{X}_{\hat{\theta}} - \theta)^2 \tag{1.3.7}$$

Suponha-se que $\hat{\theta}_n$ é um estimador de um parâmetro θ baseado em n variáveis aleatórias $X_1, \dots, X_n$. À medida que n aumenta, parece razoável esperar que a distribuição amostral de $\hat{\theta}_n$ deverá tornar-se cada vez mais concentrada em torno do verdadeiro valor do parâmetro θ. Esta propriedade da sequência de estimadores $\{\hat{\theta}_n\}$ é conhecida por consistência. Uma sequência de estimadores $\{\hat{\theta}_n\}$ é dita consistente para θ se $\{\hat{\theta}_n\}$ converge em probabilidade para θ, isto é se:

$$\lim_{n\to\infty} P\left[\left|\hat{\theta}_n - \theta\right| > \varepsilon\right] = 0 \tag{1.3.8}$$

para cada $\varepsilon > 0$ e para cada valor de θ.

A condição seguinte, que assume a existência de momentos de segunda ordem, frequentemente proporciona um método conveniente para provar a consistência de um estimador. Assim uma sequência de estimadores $\{\hat{\theta}_n\}$ é dita consistente para θ se:

$$\lim_{n\to\infty} E\left[(\hat{\theta}_n - \theta)^2\right] = 0 \tag{1.3.9}$$

Pode ser demonstrado que a equação (1.3.9) é equivalente à equação (1.3.8). Para mais detalhes veja-se (Lehmann & Casella, 1998). Pelas equações (1.3.1), (1.3.3) e (1.3.9) pode concluir-se que as duas equações seguintes são condições suficientes de consistência. Devendo ambas se verificar em simultâneo para que o estimador seja consistente.

$$\lim_{n\to\infty} E\left[\hat{\theta}_n\right] = \theta \tag{1.3.10}$$

$$\lim_{n\to\infty} Var[\hat{\theta}_n] = 0 \qquad (1.3.11)$$

Dados dois estimadores de um mesmo parâmetro θ, é preferível aquele que tiver menor erro quadrático médio, dizendo-se que este é o mais eficiente. Embora em certos casos, se possa optar por um estimador enviesado, um estimador deve ser sempre consistente.

O erro quadrático médio pode ter utilidade como método de comparação entre estimadores para um mesmo parâmetro. A eficiência relativa dada pela razão entre o erro quadrático médio de dois estimadores para um mesmo parâmetro, proporciona um método de comparação entre ambos os estimadores, que facilita muito a identificação do estimador mais eficiente, ou seja aquele que resulta no menor erro quadrático médio. Assim se o erro quadrático médio do estimador relativo ao numerador for menor que o erro quadrático médio do estimador correspondente ao denominador, a eficiência relativa resulta inferior à unidade. Se o erro quadrático médio do estimador relativo ao denominador for o menor, então a eficiência relativa será superior à unidade. Caso haja igualdade do erro quadrático médio de ambos os estimadores, a eficiência relativa será exactamente igual à unidade. Estas palavras podem ser expressas em linguagem matemática da seguinte forma. Defina-se a eficiência relativa do estimador A em relação ao estimador B, ambos estimadores para um dado parâmetro θ:

$$e_{A/B} = \frac{EQM_A(\hat{\theta})}{EQM_B(\hat{\theta})} \qquad (1.3.12)$$

Então $e_{(A/B)} = 1$ implica $EQM_A(\hat{\theta}) = EQM_B(\hat{\theta})$, sendo os estimadores A e B igualmente eficientes. Caso $e_{(A/B)} < 1$ então $EQM_A(\hat{\theta}) < EQM_B(\hat{\theta})$, sendo o estimador A o mais eficiente. Quando ocorrer $e_{(A/B)} > 1$ então $EQM_B(\hat{\theta}) < EQM_A(\hat{\theta})$, pelo que o estimador B será o mais eficiente entre os dois estimadores.

1.4 Estimador Kaplan-Meier

Todos os estimadores da função de distribuição bivariada apresentados neste trabalho se baseiam em estimadores univariados, desempenhando o estimador Kaplan-Meier um papel determinante em todos eles. Será por isso feita uma breve abordagem ao estimador Kaplan-Meier. Um estudo detalhado do mesmo estimador pode ser conseguido por meio da leitura de algumas das

referências deste trabalho, nomeadamente (Kaplan & Meier, 1958), (Hosmer, Jr. & Lemeshow, 1999), (Hougaard, 2000), (Klein & Moeschberger, 2003), (Lee & Wang, 2003) e (Borgan, 2005).

Uma característica peculiar, frequentemente presente em dados que envolvem o tempo decorrido desde um instante inicial até à ocorrência de um evento de interesse, é conhecida por censura. A censura ocorre quando para alguns indivíduos em estudo, não é observada a realização do evento de interesse durante o período em que esses indivíduos estão em observação. Podendo portanto existir uma ausência parcial ou total de informação. Existem vários tipos de censura, tais como censura pela direita, censura pela esquerda e censura intervalar. Uma discussão mais pormenorizada destes tipos de censura, bem como uma modelação estatística dos mesmos, pode ser encontrada no livro de (Klein & Moeschberger, 2003, pp. 63-78). Todos os estimadores univariados e bivariados abordados neste trabalho assumem censura pela direita. Censura pela direita ocorre quando o evento de interesse apenas é observado se o mesmo ocorrer antes de um tempo predefinido, que geralmente determina o fim de um estudo estatístico.

O estimador Kaplan-Meier, também designado estimador produto-limite é um estimador da função de sobrevivência não paramétrico.[2] Seja n_i o número de indivíduos em risco de observar um evento no tempo $t_i, i = 1,2,\dots,n;$ e o número de eventos observados no mesmo tempo representado por d_i. O estimador Kaplan-Meier da função de sobrevivência definido no tempo t é obtido a partir da equação:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{1.4.1}$$

Com a convenção:

$$\hat{S}(t) = 1 \; se \; t < t_1 \tag{1.4.2}$$

Note-se a semelhança do estimador Kaplan-Meier com a função de sobrevivência teórica definida para variáveis aleatórias discretas dada pela equação (1.2.7). Para valores de t maiores que o maior tempo observado, este estimador não está bem definido. Neste trabalho assume-se que $\hat{S}(t) = \hat{S}(t_{max})$ para tempos $t > t_{max}$, o que significa assumir que este indivíduo observaria o evento num tempo ∞. Esta solução é sugerida por (Klein & Moeschberger, 2003, p. 100). O estimador produto-limite é uma função em escada com saltos nos tempos que observam evento. (Figura 1.2).

[2] O estimador Kaplan-Meier é classificado como não paramétrico, dado não assumir qualquer modelo de probabilidade paramétrico em particular.

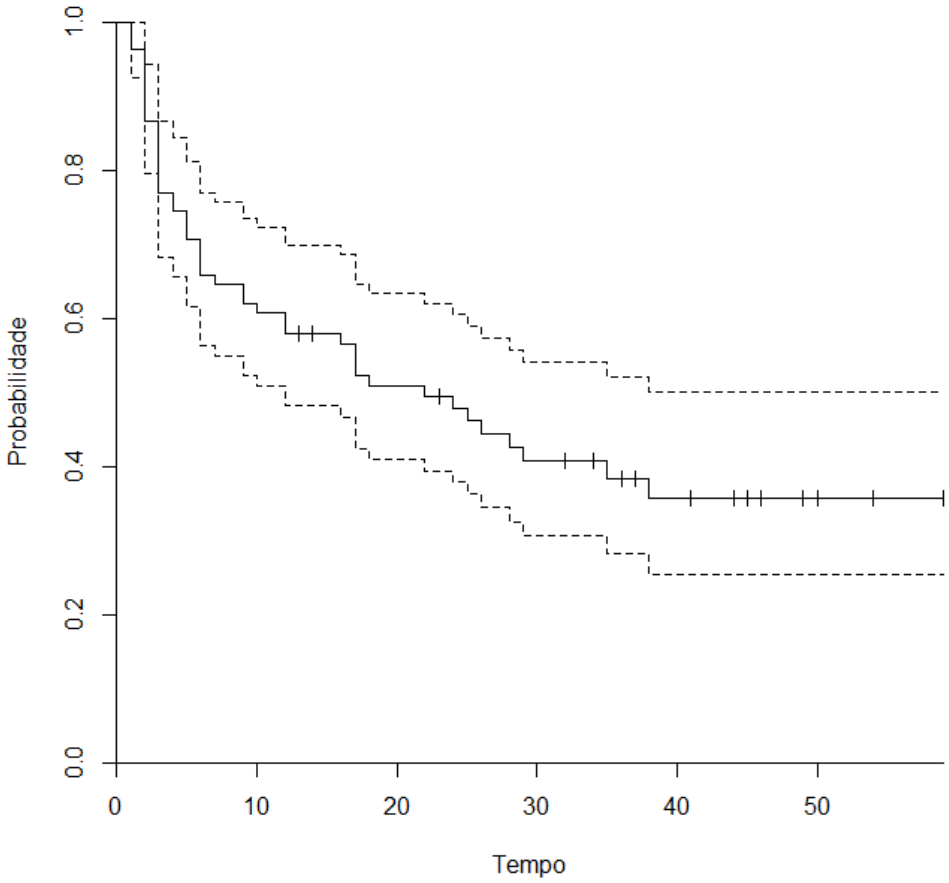


Figura 1.2: Curva de sobrevivência Kaplan-Meier com bandas de confiança a 95%.

O tamanho destes saltos depende do número de eventos observados em cada tempo t_i, bem como do padrão de observações censuradas observadas nos tempos anteriores a t_i. Em seguida o estimador Kaplan-Meier será redefinido de uma forma que facilita o seu cálculo computacional e que simplifica a derivação da expressão para a estimação dos referidos saltos, aqui designados por pesos Kaplan-Meier.

Seja T uma variável aleatória contínua que representa o tempo até um evento de interesse. Admita-se que a variável aleatória anteriormente referida é sujeita a censura aleatória pela direita. Seja C a variável aleatória de censura pela direita. Considere-se que C é independente de T. Devido à censura observa-se $(Y_i, \Delta_i), i = 1,2, \dots, n$, que são n réplicas independentes de (Y, Δ), onde $Y_i = \min(T_i, C_i)$ e $\Delta_i = I(T_i \leq C_i)$. Assim a variável Δ_i é designada por indicador de evento ou de censura, sendo igual a 1 quando for observado o evento e igual a 0 quando o evento for censurado. Ordenando os tempos $t_1 < t_2 < \cdots < t_n$, por ordem crescente dos mesmos, onde para tempos

iguais, os pares (t_i, Δ_i) cujos $\Delta_i = 1$ precedem os pares (t_i, Δ_i) cujos $\Delta_i = 0$. Estando a amostra ordenada desta forma, o estimador Kaplan-Meier da função de sobrevivência, pode ser escrito da seguinte forma:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{\Delta_i}{n - i + 1}\right) \tag{1.4.3}$$

sendo n a dimensão da amostra; Δ_i o indicador de evento anteriormente definido; e $i = 1,2,..,n$ o índice da observação na amostra.

O tamanho dos saltos anteriormente referido, a que se designou por pesos Kaplan-Meier, pode agora ser facilmente derivado recorrendo à equação (1.1.7) e ao estimador Kaplan-Meier sob a forma (1.4.3). Assim pode escrever-se:

$$\hat{S}(t_i) = \prod_{j=1}^{i} \left(1 - \frac{\Delta_j}{n - j + 1}\right) \tag{1.4.4}$$

$$\hat{S}(t_{i-1}) = \prod_{j=1}^{i-1} \left(1 - \frac{\Delta_j}{n - j + 1}\right) \tag{1.4.5}$$

Pelas expressões (1.4.4) e (1.4.5) pode observar-se que:

$$\hat{S}(t_i) = \left(1 - \frac{\Delta_i}{n - i + 1}\right) \hat{S}(t_{i-1}) \tag{1.4.6}$$

Assim pela equação (1.1.7) o peso Kaplan-Meier da observação índice i denotado por W_i pode assim ser determinado:

$$\begin{aligned} W_i = \hat{p}(t_i) &= \hat{S}(t_{i-1}) - \hat{S}(t_i) \\ &= \hat{S}(t_{i-1}) - \left(1 - \frac{\Delta_i}{n - i + 1}\right) \hat{S}(t_{i-1}) \\ &= \hat{S}(t_{i-1}) \left(1 - 1 + \frac{\Delta_i}{n - i + 1}\right) \end{aligned} \tag{1.4.7}$$

Voltando à equação (1.4.5) chega-se à expressão final para os pesos Kaplan-Meier:

$$W_i = \frac{\Delta_i}{n - i + 1} \prod_{j=1}^{i-1} \left(1 - \frac{\Delta_j}{n - j + 1}\right) \tag{1.4.8}$$

Note-se ainda que a função de sobrevivência pode ser estimada a partir dos pesos Kaplan-Meier, segundo a equação (1.1.6) e a notação anteriormente adoptada vem:

$$\hat{S}(t) = \sum_{i=i}^{n} W_i I(Y_i > t) \tag{1.4.9}$$

Pela equação (1.1.8) a função de distribuição pode ser estimada à custa dos pesos Kaplan-Meier:

$$\hat{F}(t) = \sum_{i=i}^{n} W_i I(Y_i \leq t) \tag{1.4.10}$$

Ou em alternativa, recorrendo às equações (1.1.2) e (1.4.9), a função de distribuição pode ser estimada:

$$\hat{F}(t) = 1 - \sum_{i=i}^{n} W_i I(Y_i > t) \tag{1.4.11}$$

Note-se que os estimadores da função de distribuição dados pelas expressões (1.4.10) e (1.4.11) apenas são iguais quando nenhuma das observações presentes na amostra for censurada, pois só nessa situação a soma de todos os pesos Kaplan-Meier será igual à unidade, com cada um dos pesos Kaplan-Meier igual a $\frac{1}{n}$.

Frequentemente há interesse na estimação das funções de distribuição e de sobrevivência da variável aleatória de censura. A função de sobrevivência da censura denotada por $G(.)$ pode ser estimada recorrendo ao estimador do produto-limite definido pela equação (1.4.1), invertendo o papel do evento em relação à censura:

$$\hat{G}(t) = \hat{P}[C > t] = \prod_{t_i \leq t} \left(1 - \frac{r_i}{n_i}\right) \tag{1.4.12}$$

Onde r_i será igual ao número de observações censuradas no tempo t_i. Seja y_i o número de observações no tempo t_i, pode escrever-se:

$$r_i = \sum_{j=1}^{y_i} (1 - \Delta_j) = \sum_{j=1}^{y_i} 1 - \sum_{j=1}^{y_i} \Delta_j = y_i - d_i \tag{1.4.13}$$

Como T é independente de C tem-se:

$$\hat{P}[\min(T, C) > t] = \hat{P}[T > t]\hat{P}[C > t] \tag{1.4.14}$$

Sendo $\min(T, C) = Y$ como anteriormente definido. A função de sobrevivência do mínimo pode também ser estimada recorrendo ao estimador Kaplan-Meier definido em (1.4.1), considerando que todas as observações são evento:

$$\hat{P}[Y > t] = \prod_{t_i \leq t} \left(1 - \frac{y_i}{n_i}\right) \tag{1.4.15}$$

Ora verifica-se que o produto dos lados direitos das equações (1.4.1) e (1.4.12) não é igual a (1.4.15), como determina (1.4.14), pelo que o pressuposto da independência entre as variáveis aleatórias T e C é violado. Assim é preferível definir um estimador para a função de sobrevivência da censura que não viole a equação (1.4.14). O seguinte estimador satisfaz esse requisito:

$$\hat{G}(t) = \hat{P}[C > t] = \prod_{t_i \leq t} \left(1 - \frac{r_i}{n_i - d_i}\right) \tag{1.4.16}$$

A demonstração é relativamente simples:

$$\hat{P}[C > t]\hat{P}[T > t] = \prod_{t_i \leq t} \left(1 - \frac{y_i - d_i}{n_i - d_i}\right) \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

$$= \prod_{t_i \leq t} \left(\frac{n_i - d_i - y_i + d_i}{n_i - d_i}\right)\left(\frac{n_i - d_i}{n_i}\right) = \prod_{t_i \leq t} \left(\frac{n_i - y_i}{n_i}\right) \tag{1.4.17}$$

$$= \prod_{t_i \leq t} \left(1 - \frac{y_i}{n_i}\right) = \hat{P}[Y > t]$$

Repare-se que quando numa amostra, para todos os tempos iguais, os respectivos indicadores de evento forem todos iguais a 1 ou todos iguais a 0, os estimadores (1.4.12) e (1.4.16) proporcionam exactamente os mesmos valores. Sabendo que a função de sobrevivência do mínimo pode também ser estimada recorrendo ao estimador empírico $\hat{P}[\mathrm{Y} > t] = \frac{1}{n}\sum_{i=1}^{n} I(Y_i > t)$, pela equação (1.4.14), o estimador Kaplan-Meier dado pela equação (1.4.1) é equivalente ao estimador seguinte:

$$\hat{S}(t) = \frac{1}{n\hat{G}(t)} \sum_{i=1}^{n} I(Y_i > t) \tag{1.4.18}$$

Onde $\hat{G}(t)$ é o estimador da função de sobrevivência da variável aleatória de censura definido em (1.4.16).

Uma estimativa para a função de sobrevivência pode ser determinada recorrendo a diversas abordagens. A seguinte abordagem permite obter um estimador equivalente ao estimador Kaplan-

Meier, particularmente relevante no contexto do presente trabalho. Assim recorrendo à notação e aos pressupostos anteriormente introduzidos, seja $S(.)$ a função de sobrevivência da variável aleatória T, e $G(.)$ a função de sobrevivência da variável aleatória C. Sabendo que $C_i \geq T_i$ implica $Y_i = \min(T_i, C_i) = T_i$ e assumindo que T_i é independente de C_i, pode ser estabelecida a seguinte relação:

$$\begin{aligned} & I(Y_i > t, C_i \geq T_i) = I(T_i > t, C_i \geq T_i) \\ \Leftrightarrow\ & I(Y_i > t)\Delta_i = I(T_i > t)I(C_i > T_i^-) \\ \Leftrightarrow\ & E[I(Y_i > t)\Delta_i] = S(t)G(T_i^-) \\ \Leftrightarrow\ & S(t) = E\left[\frac{I(Y_i > t)\Delta_i}{G(T_i^-)}\right] \end{aligned} \tag{1.4.19}$$

No terceiro passo foram aplicadas esperanças matemáticas a ambos os lados da equação. Se a função $G(.)$ for conhecida, a equação final obtida em (1.4.19) proporciona uma estimativa centrada para a função $S(.)$. Substituindo $G(.)$ por uma sua estimativa e estimando a esperança matemática pela esperança amostral, vem:

$$\hat{S}(t) = \frac{1}{n}\sum_{i=1}^{n}\frac{I(Y_i > t)\Delta_i}{\hat{G}(Y_i^-)} \tag{1.4.20}$$

Através de um processo idêntico ao utilizado para obter a equação (1.4.8), pode ser obtido o salto em cada observação de índice i, aqui denotado por K_i:

$$K_i = \frac{\Delta_\mathrm{i}}{n\hat{G}(Y_i^-)} \tag{1.4.21}$$

(Satten & Datta, 2001, p. 208) demonstram que quando $\hat{G}(.)$ é igual ao estimador produto-limite da censura dado pela equação (1.4.16), a equação (1.4.21) é equivalente à equação (1.4.8), estabelecendo assim a equivalência entre o estimador (1.4.20) e o estimador Kaplan-Meier dado pela expressão (1.4.3).

Sob determinadas condições, pode ser demonstrado que o estimador de Kaplan-Meier é um estimador de máxima verosimilhança não paramétrico. Pode ainda ser demonstrado que o mesmo estimador é centrado e consistente. As referidas demonstrações podem ser consultadas no artigo de (Kaplan & Meier, 1958).

Os intervalos de confiança ou bandas de confiança são muito úteis no desenvolvimento de inferências em relação a um estimador. A variância de um estimador é indispensável à construção de intervalos ou bandas de confiança. Diversos estimadores para a variância do estimador Kaplan-Meier foram propostos na literatura, sendo o estimador de Greenwood o mais usado:

$$\widehat{Var}\left[\hat{S}(t)\right] = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)} \tag{1.4.22}$$

A derivação detalhada deste estimador recorrendo ao método delta, pode ser encontrada na obra de (Hosmer, Jr. & Lemeshow, 1999, pp. 354-357). Intervalos de confiança e bandas de confiança para o estimador de Kaplan-Meier podem ser consultados se necessário em (Klein & Moeschberger, 2003, pp. 104-117).

1.5 Distribuições bivariadas

Serão introduzidos alguns conceitos de Estatística Multivariada de extrema relevância. Para o presente trabalho interessam apenas os modelos multivariados com duas variáveis aleatórias, pelo que não será apresentada a extensão para um número ilimitado de variáveis aleatórias. Esta abordagem simplifica muito as expressões, não havendo necessidade de recorrer à teoria da álgebra linear, que envolve vectores e matrizes. As obras de (Hardle & Simar, 2007), (Press, 2005) e (Reis, 2001) foram consultadas e são recomendadas caso o leitor pretenda uma abordagem propriamente multivariada à referida teoria.

Sejam X e Y duas variáveis aleatórias definidas conjuntamente, ou seja, as variáveis aleatórias X e Y têm função de distribuição cumulativa conjunta:

$$F(x, y) = P[X \leq x, Y \leq y] \tag{1.5.1}$$

De acordo com (Press, 2005, p. 56), toda a função de distribuição cumulativa bivariada[3] $F(x, y)$ satisfaz as propriedades seguintes:

[3] Por uma questão de simplicidade será de agora em diante designada por função de distribuição bivariada.

(1) $F(x,y)$ é monótona não decrescente em cada componente x e y,
(2) $0 \leq F(x,y) \leq 1$,
(3) $F(-\infty,y) = F(x,-\infty) = 0$,
(4) $F(+\infty,+\infty) = 1$,
(5) A probabilidade de qualquer rectângulo bidimensional é não negativa, assim para $x_1 < x_2, y_1 < y_2$:

$$P[rectângulo] = P[x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2]$$
$$= F(x_2,y_2) - F(x_2,y_1) - F(x_1,y_2) + F(x_1,y_1) \geq 0. \quad (1.5.2)$$

Seja $f(u,v)$ uma função densidade de probabilidade bivariada contínua, então:

$$F(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f(u,v)du\,dv \quad (1.5.3)$$

Pelo que para variáveis aleatórias contínuas:

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y} \quad (1.5.4)$$

Para variáveis aleatórias discretas, dada a função massa de probabilidade bivariada $f(x_i,y_i) = P[X = x_i, Y = y_i]$, a função de distribuição bivariada define-se por:

$$F(x,y) = \sum_{x_i \leq x}\sum_{y_i \leq y} f(x_i,y_i) \quad (1.5.5)$$

Uma função densidade de probabilidade ou função de distribuição cumulativa calculada para uma só variável a partir de uma função conjunta toma o nome de função densidade de probabilidade marginal ou distribuição cumulativa marginal, respectivamente. Suponha-se que é dada a função densidade de probabilidade conjunta $f(x,y)$ e se pretende obter as funções densidade de probabilidade marginais $g(x)$ e $h(y)$:

$$g(x) = \int_{-\infty}^{+\infty} f(x,y)dy$$
$$h(y) = \int_{-\infty}^{+\infty} f(x,y)dx \quad (1.5.6)$$

Para variáveis aleatórias discretas pode escrever-se:

$$g(x) = \sum_{y_i} f(x, y_i)$$
$$h(y) = \sum_{x_i} f(x_i, y) \tag{1.5.7}$$

Caso seja conhecida a função de distribuição cumulativa conjunta $F(x, y)$ e se pretenda obter as funções de distribuição cumulativas marginais $G(x)$ e $H(y)$:

$$G(x) = F(x, +\infty)$$
$$H(y) = F(+\infty, y) \tag{1.5.8}$$

Sendo as expressões (1.5.8) válidas tanto para variáveis aleatórias discretas como para variáveis aleatórias contínuas.

As equações (1.5.3) e (1.5.5) que relacionam a função densidade ou massa de probabilidade bivariada com a função de distribuição bivariada, têm uma interpretação geométrica que pode ser muito útil, na determinação de equações que relacionam a função de distribuição bivariada com as funções de distribuição univariadas. Estas equações, por sua vez podem ser úteis no desenvolvimento de estimadores para a função de distribuição bivariada com as propriedades estatísticas desejáveis referidas no ponto 1.3. (Lin, Sun, & Ying, 1999) sugerem um estimador obtido à custa deste princípio. Considere-se um referencial cartesiano tridimensional com eixos x, y e z. Considere-se ainda uma função densidade de probabilidade $f(x, y)$ e a respectiva função de distribuição $F(t1, t2)$. A função densidade de probabilidade bivariada representada num referencial cartesiano tridimensional será uma superfície, cuja forma depende dessa mesma função $f(x, y)$. Pois cada ponto P dessa mesma superfície terá coordenadas $P \rightarrow \big(x, y, f(x, y)\big)$. A função de distribuição $F(t1, t2)$ é igual ao volume limitado pela superfície dada pela função densidade de probabilidade $f(x, y)$, pelo plano $z = 0$, pelo plano $x = t1$ e pelo plano $y = t2$. Pode notar-se que existe a seguinte relação entre conjuntos de pontos:

$$\{x \leq t1 \wedge z = 0\}$$
$$= \{x \leq t1 \wedge y \leq t2 \wedge z = 0\} \cup \{x \leq t1 \wedge y > t2 \wedge z = 0\} \tag{1.5.9}$$

De modo a auxiliar este raciocínio foram construídos os gráficos da Figura 1.3.

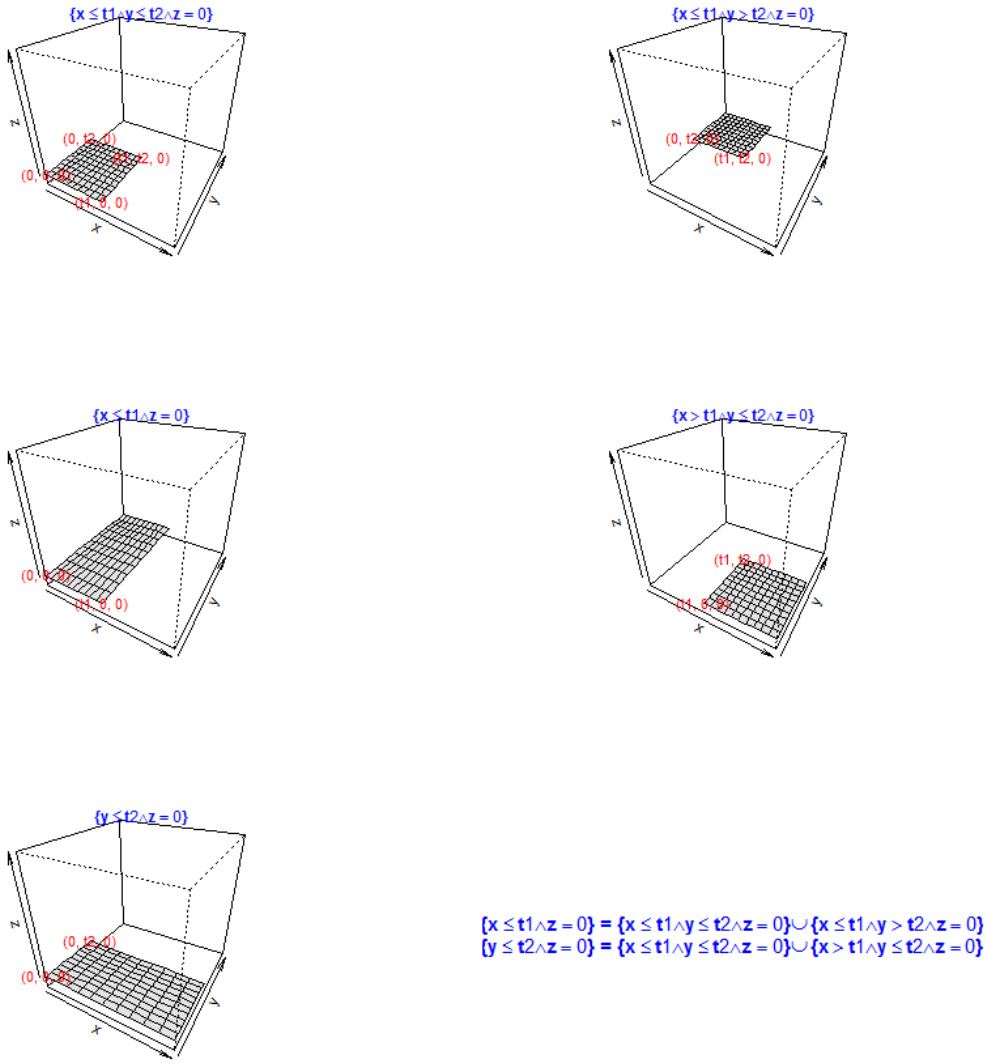


Figura 1.3: Representações gráficas de conjuntos em referenciais cartesianos tridimensionais.

O mesmo gráfico sugere ainda a existência da seguinte relação:

$$\{y \leq t2 \wedge z = 0\}$$
$$= \{x \leq t1 \wedge y \leq t2 \wedge z = 0\} \cup \{x > t1 \wedge y \leq t2 \wedge z = 0\} \quad (1.5.10)$$

A intuição leva a crer que existe uma relação idêntica entre os volumes limitados pelos conjuntos de pontos representados nos diversos gráficos da Figura 1.3, e o conjunto de pontos dado por uma função qualquer $f(x, y)$. Assim de acordo com a interpretação geométrica que relaciona uma

função de distribuição $F(t1,t2)$ com a função densidade de probabilidade associada $f(x,y)$, conclui-se o seguinte:

$$F(t1,t2) = F(t1) - P[x \leq t1 \wedge y > t2]$$
$$F(t1,t2) = F(t2) - P[x > t1 \wedge y \leq t2] \quad (1.5.11)$$

Onde $F(t1)$ e $F(t2)$ são as funções de distribuição marginais de $F(t1,t2)$ e $P[.]$ representa a probabilidade do conjunto apresentado no interior do parêntesis recto. Seguindo um raciocínio idêntico, é possível chegar à expressão para o cálculo da probabilidade do rectângulo referida na propriedade (5) em (1.5.2).

O modelo de probabilidade Weibull constitui um modelo paramétrico bastante versátil para modelar fenómenos que envolvem o tempo até um evento de interesse. O modelo de probabilidade Weibull univariado tem função de sobrevivência dada pela expressão:

$$S(x) = P[X > x] = exp\left\{-\left(\frac{x}{\theta}\right)^{\beta}\right\}, \quad (1.5.12)$$
$$\theta > 0, \beta > 0, x \geq 0$$

onde θ é designado por parâmetro de escala e β é designado por parâmetro de forma. Uma função de distribuição bivariada Weibull, para além de verificar as propriedades enunciadas no ponto (1.5.2), tem funções de distribuição marginais Weibull. Pelo que as funções de sobrevivência marginais deverão ter a forma (1.5.12). É possível definir várias funções de distribuição bivariadas cujas funções de distribuição marginais são do tipo Weibull. Várias funções de distribuição Weibull bivariadas não equivalentes entre si foram propostas na literatura. (Lu & Bhattacharyya, 1990) propuseram um modelo de probabilidade Weibull bivariado baseado em riscos aleatórios, que foi adoptado neste trabalho. Dadas X e Y duas variáveis aleatórias, o modelo de probabilidade Weibull bivariado referido, tem função de sobrevivência:

$$S(x,y) = P[X > x, Y > y] = exp\left\{-\left[\left(\frac{x}{\theta_1}\right)^{\frac{\beta_1}{\delta}} + \left(\frac{y}{\theta_2}\right)^{\frac{\beta_2}{\delta}}\right]^{\delta}\right\},$$
$$\theta_1 > 0, \beta_1 > 0, x \geq 0, \quad (1.5.13)$$
$$\theta_2 > 0, \beta_2 > 0, y \geq 0,$$
$$0 < \delta \leq 1$$

Facilmente se verifica que as funções de sobrevivência marginais são Weibull, sendo obtidas expressões idênticas a (1.5.12):

$$S(x) = S(x, -\infty) = S(x, 0) = exp\left\{-\left(\frac{x}{\theta_1}\right)^{\beta_1}\right\}$$
$$S(y) = S(-\infty, y) = S(0, y) = exp\left\{-\left(\frac{y}{\theta_2}\right)^{\beta_2}\right\} \quad (1.5.14)$$

Verifica-se que quando $\delta = 1$, as variáveis aleatórias são independentes entre si, pois nesse caso $S(x, y) = S(x)S(y)$. A covariância entre as variáveis aleatórias X e Y $Cov[X, Y]$, é uma função complexa dos parâmetros do modelo bivariado $\theta_1, \beta_1, \theta_2, \beta_2$ e δ, veja-se (Lu & Bhattacharyya, 1990, p. 555) ou (Johnson, Evans, & Green, 1999, p. 3). A função densidade de probabilidade bivariada é uma expressão relativamente longa, pelo que não será aqui referida. A mesma função consta no artigo de (Lu & Bhattacharyya, 1990, p. 554), bem como no artigo de (Johnson, Evans, & Green, 1999, p. 5). Uma expressão para a função de distribuição bivariada não é referida em nenhuma das referências bibliográficas deste trabalho. No entanto pode ser determinada a partir da função densidade de probabilidade bivariada de acordo com a equação (1.5.3). Tendo em vista a complexidade da função densidade de probabilidade bivariada, a integração dupla da mesma pode ser uma tarefa demasiado demorada, pelo que esta via não foi sequer considerada. Em alternativa a função de distribuição bivariada pode ser determinada recorrendo a um método mais expedito. Com o auxílio de um esboço gráfico idêntico ao da Figura 1.3, pode-se concluir que a função de distribuição bivariada relaciona-se com as funções de sobrevivência marginais e com a função de sobrevivência bivariada de acordo com a expressão:

$$F(x, y) = 1 - S(x) - S(y) + S(x, y) \quad (1.5.15)$$

Pelo que a função de distribuição bivariada Weibull pode escrever-se:

$$F(x, y) = P[X \leq x, Y \leq y]$$
$$= 1 - exp\left\{-\left(\frac{x}{\theta_1}\right)^{\beta_1}\right\} - exp\left\{-\left(\frac{y}{\theta_2}\right)^{\beta_2}\right\} + exp\left\{-\left[\left(\frac{x}{\theta_1}\right)^{\frac{\beta_1}{\delta}} + \left(\frac{y}{\theta_2}\right)^{\frac{\beta_2}{\delta}}\right]^{\delta}\right\} \quad (1.5.16)$$

Obviamente que os parâmetros bem como as variáveis se encontram dentro dos limites especificados em (1.5.13).

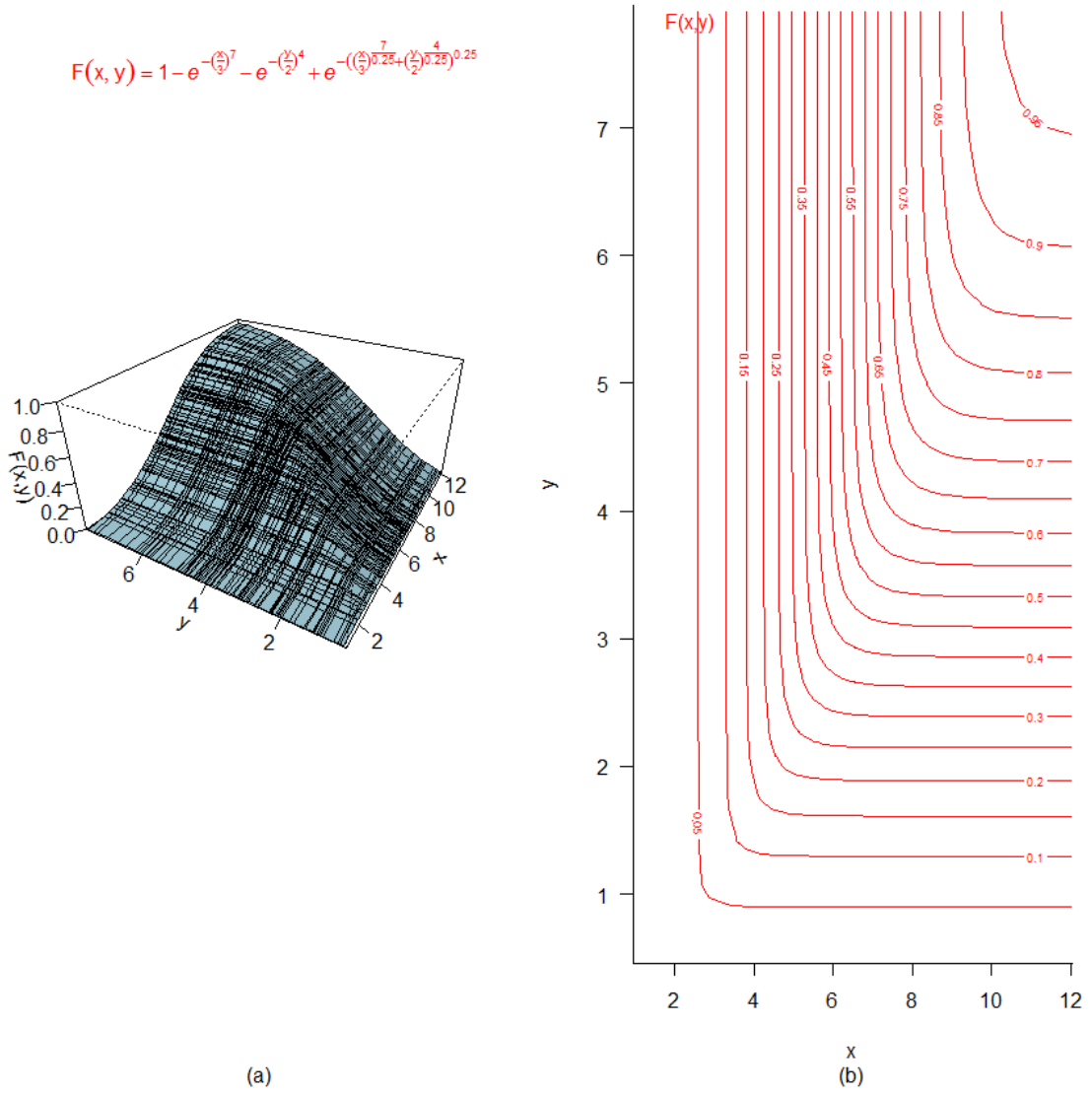


Figura 1.4: Representações gráficas de uma função de distribuição bivariada Weibull. (a) Perspectiva. (b) Linhas de nível.

O gráfico da Figura 1.4 permite visualizar representações gráficas de uma função de distribuição bivariada Weibull com parâmetros $\theta_1 = 3, \beta_1 = 7, \theta_2 = 2, \beta_2 = 4, \delta = 0.25$.

Neste trabalho foi ainda adoptado o modelo de probabilidade bivariado da família de Farlie-Gumbel-Morgenstern com marginais exponenciais. Segundo (Kotz, Balakrishnan, & Johnson, 2000, p. 52) a função de distribuição da família Farlie-Gumbel-Morgenstern é dada pela equação:

$$F(x,y) = P[X \leq x, Y \leq y] \\ = F(x)F(y)[1 + \alpha\{1 - F(x)\}\{1 - F(y)\}], |\alpha| \leq 1 \quad (1.5.17)$$

onde $F(x)$ e $F(y)$ são as funções de distribuição marginais. Quando $\alpha = 0$ as variáveis aleatórias são independentes, pois nesse caso $F(x,y) = F(x)F(y)$.

A função de distribuição univariada exponencial pode ser obtida a partir da função de distribuição Weibull univariada (1.5.12) com parâmetro de forma igual à unidade:

$$F(x) = P[X \leq x] = 1 - exp\left\{-\frac{x}{\theta}\right\}, \quad \theta > 0, x \geq 0 \tag{1.5.18}$$

Assim a função de distribuição bivariada exponencial de Farlie-Gumbel-Morgenstern pode escrever-se:

$$F(x,y) = P[X \leq x, Y \leq y]$$
$$= \left(1 - exp\left\{-\frac{x}{\theta_1}\right\}\right)\left(1 - exp\left\{-\frac{y}{\theta_2}\right\}\right)\left(1 + \alpha exp\left\{-\left(\frac{x}{\theta_1} + \frac{y}{\theta_2}\right)\right\}\right), \tag{1.5.19}$$
$$\theta_1 > 0, x \geq 0,$$
$$\theta_2 > 0, y \geq 0,$$
$$-1 \leq \alpha \leq 1$$

A função densidade de probabilidade dos modelos de probabilidade da família de Farlie-Gumbel-Morgenstern obtém-se derivando (1.5.17) em ordem a cada uma das variáveis segundo a equação (1.5.4):

$$f(x,y) = f(x)f(y)[1 + \alpha\{1 - 2F(x)\}\{1 - 2F(y)\}], |\alpha| \leq 1 \tag{1.5.20}$$

Para funções densidade de probabilidade marginais ambas exponenciais, tem-se:

$$f(x,y) = f_{\theta_1}(x)f_{\theta_2}(y) + \alpha\left\{f_{\frac{\theta_1}{2}}(x) - f_{\theta_1}(x)\right\}\left\{f_{\frac{\theta_2}{2}}(y) - f_{\theta_2}(y)\right\}, \tag{1.5.21}$$
$$\theta_1 > 0, x \geq 0,$$
$$\theta_2 > 0, y \geq 0,$$
$$-1 \leq \alpha \leq 1$$

onde $f_{\theta_1}(x)$ denota a função densidade de probabilidade univariada exponencial de parâmetro θ_1, e $f_{\frac{\theta_1}{2}}(x)$ a função densidade de probabilidade univariada exponencial de parâmetro $\frac{\theta_1}{2}$. Sendo seguido

o mesmo princípio em relação a $f_{\theta_2}(y)$ e $f_{\frac{\theta_2}{2}}(y)$. A função densidade de probabilidade bivariada exponencial sob esta forma facilita a obtenção da covariância entre ambas as variáveis aleatórias:

$$Cov[X,Y] = \frac{1}{4}\alpha\theta_1\theta_2 \tag{1.5.22}$$

Pelo que para o modelo de probabilidade bivariado de Farlie-Gumbel-Morgenstern com marginais exponenciais a correlação depende apenas do parâmetro α:

$$Corr[X,Y] = \frac{1}{4}\alpha \tag{1.5.23}$$

Como $|\alpha| \leq 1$, a correlação não pode exceder $\frac{1}{4}$ ou ser inferior a $-\frac{1}{4}$. O gráfico da Figura 1.5 permite visualizar representações gráficas de uma função de distribuição bivariada exponencial da família Farlie-Gumbel-Morgenstern com parâmetros $\theta_1 = 1, \theta_2 = 1, \alpha = -1$.

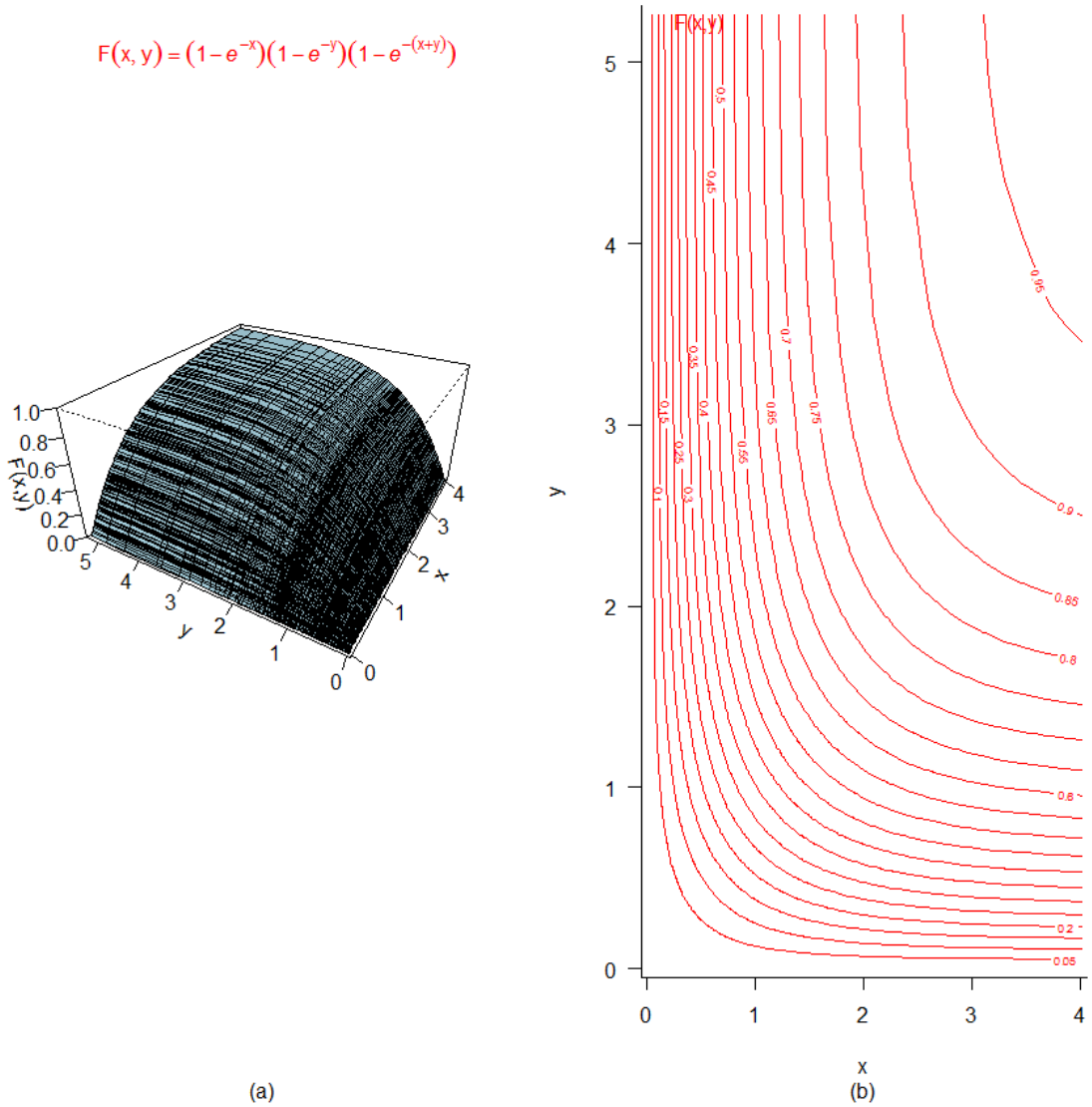


Figura 1.5: Representações gráficas de uma função de distribuição bivariada exponencial FGM. (a) Perspectiva. (b) Linhas de nível.

2 Estimação da função de distribuição bivariada para tempos sequenciais com censura pela direita

A análise estatística de intervalos de tempo consecutivos ou sequenciais é um assunto da maior importância num variado número de áreas do conhecimento, incluindo engenharia, economia, epidemiologia e análise de sobrevivência. Na maioria das vezes interessa a descrição das funções de distribuição marginais dos intervalos de tempo, bem como a estrutura de correlação entre os mesmos. O que acontece, por exemplo, quando são analisados eventos recorrentes, que surgem quando cada indivíduo pode passar por um evento bem definido várias vezes ao longo da sua história. Então, os tempos entre eventos consecutivos são referidos por intervalos de tempo, que são obviamente determinados pelos tempos nos quais os eventos recorrentes ocorrem. Exemplos de tempos sequenciais incluem as recorrências de uma determinada doença, tais como episódios de recorrências em doentes de cancro, saudável → 1ª recorrência → 2ª recorrência → …

Seja (T_1, T_2) um par de tempos obtidos a partir de eventos sequenciais, os quais são sujeitos a censura aleatória pela direita. Seja C a variável aleatória de censura pela direita, assumida independente do par (T_1, T_2). Seja $Z = T_1 + T_2$ o tempo total. Devido à censura apenas se observa $(Y_{1i}, Y_{2i}, \Delta_{1i}, \Delta_{2i}), 1 \leq i \leq n$, que são n réplicas independentes de $(Y_1, Y_2, \Delta_1, \Delta_2)$, onde $Y_1 = \min(T_1, C)$, $\Delta_1 = I(T_1 \leq C)$, $Y_2 = \min(T_2, C_2)$, $\Delta_2 = I(T_2 \leq C_2)$, sendo $C_2 = (C - T_1)\Delta_1$, a variável aleatória de censura do segundo intervalo de tempo. Defina-se ainda $\tilde{Z} = \min(Z, C)$ e seja $S_1(.)$ a função de sobrevivência de T_1, e $G(.)$ a função de sobrevivência da variável aleatória de censura C. Note-se que quando T_1 for censurado, T_2 não é observado, pelo que Δ_1 igual a zero implica Δ_2 igual a zero.

Como C e T_1 são independentes, o estimador Kaplan-Meier baseado nos pares (Y_{1i}, Δ_{1i}) pode ser utilizado para estimar consistentemente a função de sobrevivência $S_1(.)$. Similarmente, a função de sobrevivência do tempo total Z pode ser estimada consistentemente pelo estimador Kaplan-Meier baseado nos pares $(Y_{1i} + Y_{2i}, \Delta_{2i})$. Como T_2 e C_2 são geralmente dependentes, a estimação da função de distribuição marginal ou da função de sobrevivência marginal de T_2, não constitui um problema simples. O mesmo se aplica para a função de distribuição bivariada $F_{12}(x, y) = P(T_1 \leq x, T_2 \leq y)$.

Nesta secção serão apresentados quatro métodos distintos para estimar a função de distribuição bivariada $F_{12}(x, y) = P(T_1 \leq x, T_2 \leq y)$, todos baseados no estimador Kaplan-Meier da função de sobrevivência.

2.1 Estimador Kaplan-Meier condicional

Um estimador simples para a função de distribuição bivariada dos tempos sequenciais é baseado no teorema de Bayes e no estimador Kaplan-Meier para a função de sobrevivência. Sugere-se (Montgomery & Runger, 2011, pp. 55-56) ou (Pestana & Velosa, 2008, pp. 232-236) para uma introdução à probabilidade condicionada e ao teorema de Bayes. Como $F_{12}(x,y) = P(T_1 \leq x, T_2 \leq y) = P(T_2 \leq y | T_1 \leq x)P(T_1 \leq x)$, um estimador simples para a função de distribuição bivariada é dado por:

$$\hat{F}_{12}(x,y) = [1 - \hat{S}_1(x)][1 - \hat{S}_2(y|T_1 \leq x)] \qquad (2.1.1)$$

Onde $\hat{S}_1(x)$ é o estimador Kaplan-Meier baseado nos pares (Y_{1i}, Δ_{1i}) e $\hat{S}_2(y|T_1 \leq x)$ é o estimador Kaplan-Meier baseado nos pares (Y_{2i}, Δ_{2i}) para os quais $T_{1i} \leq x$. Em (Moreira & Meira-Machado, 2012, p. 3) o estimador Kaplan-Meier do segundo intervalo de tempo, baseado no par (Y_{2i}, Δ_{2i}) é condicionado ao conjunto $T_{1i} \leq x$ e $\Delta_{1i} = 1$. Na verdade não faz diferença condicionar ou não a $\Delta_{1i} = 1$, sendo ambos os estimadores equivalentes.[4] Pelo que aqui opta-se por não condicionar o estimador Kaplan-Meier do segundo intervalo de tempo ao referido conjunto.

2.2 Estimador Kaplan-Meier pesado

Outro estimador simples foi recentemente proposto por (de Uña-Álvarez & Meira-Machado, 2008). Este estimador utiliza o estimador Kaplan-Meier do tempo total para pesar os dados bivariados. O estimador proposto é dado por:

$$\hat{F}_{12}(x,y) = \sum_{i=1}^{n} W_i I(Y_{1i} \leq x, Y_{2i} \leq y) \qquad (2.2.1)$$

Onde $W_i = \frac{\Delta_{2i}}{n-i+1} \prod_{j=1}^{i-1} \left[1 - \frac{\Delta_{2j}}{n-j+1}\right]$ é o peso Kaplan-Meier associado a $\tilde{Z}_i$ quando se estima a função de distribuição marginal de Z a partir dos pares $(\tilde{Z}_i, \Delta_{2i})$. Sendo que os índices dos tempos $\tilde{Z}_i$ censurados são superiores aos índices dos tempos $\tilde{Z}_i$ não censurados caso haja empates.

[4] Neste caso sempre que $\Delta_{1i} = 0$, $Y_{2i} = 0$ e $\Delta_{2i} = 0$. Como os valores $Y_{2i} = 0$ para os quais $\Delta_{2i} = 0$ não têm qualquer influência no estimador Kaplan-Meier baseado nos pares (Y_{2i}, Δ_{2i}), é indiferente condicionar ou não o mesmo estimador ao conjunto $\Delta_{1i} = 1$.

2.3 Estimador Kaplan-Meier pesado pré-suavizado

Recentemente, (de Uña-Álvarez & Amorim, 2011) propuseram uma modificação do estimador (2.2.1) baseando-se em pré-suavização (Dikta, 1998), o qual permite uma redução de variância na presença de censura. Este estimador usa uma versão pré-suavizada do estimador Kaplan-Meier do tempo total para pesar os dados bivariados. Este estimador é expresso como:

$$\hat{F}_{12}(x,y) = \sum_{i=1}^{n} W_i^* I(Y_{1i} \leq x, Y_{2i} \leq y) \qquad (2.3.1)$$

Onde $W_i^* = \frac{m(Y_{1i},Y_{2i})}{n-i+1} \prod_{j=1}^{i-1} \left[1 - \frac{m(Y_{1j},Y_{2j})}{n-j+1}\right]$ é o peso Kaplan-Meier pré-suavizado associado a $\tilde{Z}_i$. Aqui $m(x,y) = P(\Delta_2 = 1 | Y_1 = x, Y_2 = y, \Delta_1 = 1)$ pertence a uma família de curvas de regressão binárias, por exemplo a logística. Na prática assume-se que $m(x,y) = m(x,y,\beta)$ onde β é um vector de parâmetros tipicamente determinados maximizando a verosimilhança condicional dos valores Δ_2 dado o par (Y_1, Y_2) para o qual $\Delta_1 = 1$. Na prática tem-se:

$$m(x,y,\beta) = \begin{cases} 0 & se\ x = y \\ m(x,y,\beta) & se\ x < y \end{cases} \qquad (2.3.2)$$

Neste trabalho foi adoptado o modelo de regressão logístico. Existem modelos de regressão binária alternativos que podem ser aplicados neste caso, por exemplo modelos de regressão aditivos generalizados (Wood, 2006). Um modelo de regressão logístico é um caso particular de um modelo linear generalizado onde a variável resposta neste caso Δ_2 é considerada uma variável aleatória cujo modelo de distribuição é binomial, sendo a função de ligação a função logit. A teoria de modelos de regressão linear generalizados é relativamente extensa, não podendo por isso ser aqui abordada. Existe uma vasta literatura dedicada aos modelos de regressão linear generalizados e em particular à regressão logística. Ao leitor interessado mas não familiarizado com a mesma teoria, sugere-se a consulta das obras de (Dobson & Barnett, 2008) ou (McCullagh & Nelder, 1989).

2.4 Estimador de Lin

Outro estimador para a função de distribuição bivariada foi proposto por (Lin, Sun, & Ying, 1999). O estimador de Lin é expresso por:

$$\hat{F}_{12}(x,y) = \hat{H}(x,0) - \hat{H}(x,y) \tag{2.4.1}$$

Onde:

$$\hat{H}(x,y) = \frac{1}{n}\sum_{i=1}^{n} \frac{I(Y_{1i} \leq x, Y_{2i} > y)}{\hat{G}(Y_{1i} + y)} \tag{2.4.2}$$

Sendo (2.4.1) a versão estimador de (1.5.11). A derivação detalhada do estimador (2.4.2) encontra-se em (Lin, Sun, & Ying, 1999, p. 61). A função de sobrevivência da censura $G(.)$ é tipicamente desconhecida, havendo a necessidade de a substituir por uma estimativa. Esta estimativa pode ser obtida recorrendo ao estimador Kaplan-Meier da censura na versão (1.4.16). Assim para o primeiro termo do lado direito da equação (2.4.1) pode ser utilizado o estimador (1.4.16) baseado no par (Y_{1i}, Δ_{1i}). E para o segundo termo do lado direito da equação (2.4.1) pode ser utilizado o estimador (1.4.16) baseado no par $(\tilde{Z}_i, \Delta_{2i})$. Assim o estimador (2.4.1) pode ser escrito:

$$\hat{F}_{12}(x,y) = \frac{1}{n}\sum_{i=1}^{n} \frac{I(Y_{1i} \leq x)\Delta_{1i}}{\hat{G}(Y_{1i})} - \frac{1}{n}\sum_{i=1}^{n} \frac{I(Y_{1i} \leq x, Y_{2i} > y)}{\hat{G}(Y_{1i} + y)} \tag{2.4.3}$$

Pois $\Delta_{1i} = 0$ implica $Y_{2i} = 0$, e $\Delta_{1i} = 1$ implica $Y_{2i} > 0$, pelo que $I(Y_{2i} > 0) = \Delta_{1i}$.

3 Desenvolvimento e utilização de *software*

De forma a proporcionar resultados numéricos e gráficos em relação aos estimadores da função de distribuição bivariada anteriormente abordados, foi criada uma extensão para o *software* estatístico R (R Core Team, R: A Language and Environment for Statistical Computing, 2012). A extensão à qual foi dado o nome survivalBIV (Moreira, Araújo, & Machado, 2012) foi inicialmente publicada a 26 de Outubro de 2011, tendo a esta data sido escrita exclusivamente na linguagem de programação R (R Core Team, R Language Definition, 2012). Deste então a extensão survivalBIV foi alvo de várias revisões, tendo sido introduzidos alguns melhoramentos ao nível da eficiência computacional da parte numérica. Ao nível gráfico foram introduzidas novas funcionalidades, tendo sido criadas novas funções que permitem representações gráficas bidimensionais dos resultados numéricos. A escrita de parte do código na linguagem de programação C permitiu à mesma máquina terminar em poucos segundos o que anteriormente demorava alguns minutos a concluir. O desenvolvimento da referida extensão não teria sido possível sem a ajuda de alguma literatura técnica do domínio da ciência computacional, alguma da qual se encontra disponível gratuitamente na internet. Em relação à linguagem de programação R destacam-se as obras de (Braun & Murdoch, 2007), (Crawley, 2007), (Mittal, 2011) e (Murrell, 2006), bem como o documento (R Core Team, R Language Definition, 2012). Relativamente à criação de extensões para o *software* R foram consultados os documentos (R Core Team, R Internals, 2012) e (R Core Team, Writing R Extensions, 2012). Relativamente à utilização do próprio *software* R, foram inicialmente consultados os documentos (R Core Team, R Installation and Administration, 2012) e (Venables, Smith, & R Core Team, 2012). Quanto à linguagem de programação C, o livro de (Jones & Aitken, 2003) foi estudado, tendo ocasionalmente sido consultado o livro de (Kochan, 2005). A implementação de computação paralela[5] recorrendo à "Interface de programação de aplicação"[6] OpenMP foi conseguida através da consulta da obra de (Chapman, Jost, & van der Pas, 2008), bem como dos documentos (OpenMP Architecture Review Board, OpenMP Application Program Interface, 2008) e (OpenMP Architecture Review Board, Summary of OpenMP 3.0 C/C++ Syntax, 2008).

Na data de escrita deste trabalho, a versão 1.4 é a mais recente iteração da extensão survivalBIV, não estando ainda publicada. Nesta secção será feita uma descrição da referida extensão, sendo apresentados alguns exemplos de utilização. Para este fim será utilizada a versão 1.4 da mesma extensão. De forma a reproduzir os exemplos, é necessária a instalação prévia do

[5] Forma de computação na qual muitos cálculos são realizados em simultâneo, frequentemente em unidades de processamento central multi-núcleo.
[6] Tradução do inglês "Application programming interface".

software R, bem como da versão 1.4 da extensão survivalBIV. Serão necessários conhecimentos básicos de utilização do *software* R, que não serão aqui abordados. Uma descrição detalhada dos argumentos de cada uma das funções não será apresentada, sugerindo-se a consulta do manual que acompanha a extensão survivalBIV (Moreira, Araújo, & Machado, 2012). Uma vez no ambiente do *software* R, a página de ajuda de uma qualquer função pode ser acedida recorrendo à função "help":

```
> help(bivCKM)
```

Ou adicionando o prefixo "?" ao nome da função:

```
> ?bivCKM
```

Após a execução dos referidos comandos, tipicamente abre-se uma janela independente onde consta a página de ajuda da função. Aí encontra-se uma descrição detalhada dos argumentos aceites pela função, para além de outra informação relativa à mesma função.

3.1 Utilização numérica

De forma a carregar a extensão survivalBIV é necessário correr o comando:

```
> library(survivalBIV)
```

A lista de todos os objectos visíveis para o utilizador, presentes na extensão survivalBIV pode ser assim obtida:

```
> objects("package:survivalBIV")
```

```
[1]  "bivBIV"          "bivCKM"      "bivKMPW"     bivKMW"
[5]  "bivLIN"          "bladderBIV"  "contour.BIV" "corrBIV"
[9]  "dgpBIV"          "image.BIV"   "is.survBIV"  "lines.BIV"
[13] "persp.BIV"       "plot.BIV"    "probBIV"     "summary.BIV"
[17] "summary.survBIV" "survBIV"
```

Neste caso todos os objectos são funções, excepto o objecto "bladderBIV" que é um objecto do tipo *data.frame*.[7] A base de dados "bladderBIV" será utilizada em todos os exemplos seguintes. Esta base de dados resulta de um estudo de recorrências de cancro da bexiga. Neste estudo os pacientes tinham tumores da bexiga superficiais que foram removidos por cirurgia. Muitos pacientes tiveram

[7] Em Português é comum um objecto do tipo *data.frame* ser referido por "base de dados".

múltiplas recorrências (até um máximo de 9) de tumores durante o estudo, e novos tumores foram removidos a cada consulta médica. Apenas as duas primeiras recorrências e os correspondentes intervalos de tempo (em meses) estão contempladas na base de dados. A base de dados tem 85 observações, para cada uma das 4 variáveis. Cada linha da base de dados corresponde a um indivíduo. As variáveis são; time1, tempo em meses desde que o indivíduo entrou em observação até à primeira recorrência ou censura; event1, indicador da primeira recorrência (1 se observada recorrência, 0 caso contrário); time2, intervalo de tempo em meses desde a primeira recorrência até à segunda; event2, indicador da segunda recorrência.

Por vezes é útil observar apenas parte de uma base de dados, o que pode ser conseguido à custa do comando:

```
> head(bladderBIV, n=9)
```

	time1	event1	time2	event2
1	1	0	0	0
2	4	0	0	0
3	7	0	0	0
4	10	0	0	0
5	6	1	4	0
6	14	0	0	0
7	18	0	0	0
8	5	1	13	0
9	12	1	4	1

Neste caso podem ser observadas as primeiras 9 observações de cada variável na base de dados. Assim, os primeiros 4 indivíduos, bem como os indivíduos com os números 6 e 7, não observaram qualquer recorrência enquanto estiveram em estudo. O indivíduo número 5 observou a primeira recorrência 6 meses após ter entrado em estudo, não tendo observado a segunda recorrência no período restante em que permaneceu em estudo. O indivíduo com o número 9, por sua vez registou a primeira recorrência 12 meses após ter entrado em observação, tendo registado a segunda recorrência 4 meses mais tarde.

De forma a obter resultados numéricos em relação aos quatro estimadores anteriormente apresentados, foram criadas quatro funções, uma em relação a cada estimador. Estas quatro funções aceitam o mesmo tipo de objectos como argumento e devolvem o mesmo tipo de objecto. Ao tipo de objecto devolvido foi atribuída a classe “BIV”. Os objectos devolvidos podem

posteriormente ser introduzidos como argumento em funções gráficas, podendo assim ser obtidos gráficos de interesse em relação a cada um dos estimadores. Estas funções gráficas são métodos para objectos da classe "BIV". A relação objecto classe método é uma característica da programação orientada para objectos. Um dos paradigmas da programação permitidos pela linguagem de programação R. Exemplos de utilização gráfica serão adiante demonstrados. As quatro funções são "bivCKM" para o estimador Kaplan-Meier condicional (subsecção 2.1); "bivKMW" para o estimador Kaplan-Meier pesado (subsecção 2.2); "bivKMPW" para o estimador Kaplan-Meier pesado pré-suavisado (subsecção 2.3) e "bivLIN" para o estimador de Lin (subsecção 2.4). Estas funções não lidam directamente com as bases de dados. Antes da sua utilização é necessário criar um objecto com uma estrutura apropriada, sendo este objecto utilizado como argumento nas quatro funções referidas. A função "survBIV" cria o objecto da classe "survBIV" necessário, partindo de uma base de dados qualquer:

```
> bladderBIV_obj <- with( bladderBIV, survBIV(time1, event1, time2, event2) )
```

Agora as quatro estimativas podem ser obtidas recorrendo a cada uma das quatro funções:

```
> bivCKM(object=bladderBIV_obj, t1=5, t2=20)

Conditional-Kaplan-Meier bivariate probabilities


P(time1 <= 5, time2 <= 20) = 0.192070195903072
```

```
> bivKMW(object=bladderBIV_obj, t1=5, t2=20)

Kaplan-Meier Weighted bivariate probabilities


P(time1 <= 5, time2 <= 20) = 0.192105845504139
```

```
> bivKMPW(object=bladderBIV_obj, t1=5, t2=20)

Presmoothed Kaplan-Meier Weighted bivariate probabilities
```

```
P(time1 <= 5, time2 <= 20) = 0.189784651171347
```

```
> bivLIN(object=bladderBIV_obj, t1=5, t2=20)

Lin bivariate probabilities

P(time1 <= 5, time2 <= 20) = 0.199239652200765
```

Resultados com intervalos de confiança *bootstrap* podem ser obtidos em relação a cada um dos quatro estimadores. O *bootstrap* é uma técnica relativamente simples que permite estimar a variância ou a distribuição de uma estatística. Esta técnica pode também ser usada para construir intervalos de confiança. Trata-se de uma técnica computacionalmente intensiva. Para uma breve descrição desta técnica sugere-se (Wasserman, 2006, pp. 30-35). A obra de (Efron & Tibshirani, 1994) proporciona uma introdução acessível. Sugere-se ainda a monografia de (Efron, 1982), um dos primeiros textos publicados sobre o assunto. Um intervalo de confiança a 90 % para cada um dos estimadores pode ser obtido recorrendo aos comandos seguintes:

```
> bivCKM(object=bladderBIV_obj, t1=5, t2=20, conf=TRUE, conf.level=0.9)

Conditional-Kaplan-Meier bivariate probabilities

P(time1 <= 5, time2 <= 20) = 0.192070195903072

        5%       95%
0.1214418 0.2672421
```

```
> bivKMW(object=bladderBIV_obj, t1=5, t2=20, conf=TRUE, conf.level=0.9)

Kaplan-Meier Weighted bivariate probabilities

P(time1 <= 5, time2 <= 20) = 0.192105845504139

        5%       95%
```

```
0.1197734    0.2702640
```

```
> bivKMPW(object=bladderBIV_obj, t1=5, t2=20, conf=TRUE, conf.level=0.9)

Presmoothed Kaplan-Meier Weighted bivariate probabilities


P(time1 <= 5, time2 <= 20) = 0.189784651171347

        5%          95%
0.1240061    0.2671331
```

```
> bivLIN(object=bladderBIV_obj, t1=5, t2=20, conf=TRUE, conf.level=0.9)

Lin bivariate probabilities


P(time1 <= 5, time2 <= 20) = 0.199239652200765

        5%          95%
0.1255001    0.2800840
```

O método "summary" para objectos da classe "survBIV" permite obter resultados com ou sem intervalos de confiança *bootstrap* para todos os estimadores em simultâneo:

```
> summary(bladderBIV_obj, t1=c(5,10), t2=c(20,30), method="all")

Conditional-Kaplan-Meier bivariate probabilities


Estimate of P(time1 <= t1, time2 <= t2)

              t2 = 20      t2 = 30
t1 = 5     0.1920702    0.2296734
t1 = 10    0.2600991    0.3017609


Presmoothed Kaplan-Meier Weighted bivariate probabilities
```

Estimate of P(time1 <= t1, time2 <= t2)

	t2 = 20	t2 = 30
t1 = 5	0.1897847	0.2194880
t1 = 10	0.2570558	0.2951559

Kaplan-Meier Weighted bivariate probabilities

Estimate of P(time1 <= t1, time2 <= t2)

	t2 = 20	t2 = 30
t1 = 5	0.1921058	0.2281833
t1 = 10	0.2598186	0.2958960

Lin bivariate probabilities

Estimate of P(time1 <= t1, time2 <= t2)

	t2 = 20	t2 = 30
t1 = 5	0.1992397	0.2469242
t1 = 10	0.2637906	0.3216766

Repare-se que o objecto “bladderBIV_obj” da classe “survBIV” anteriormente guardado, foi utilizado em todas as funções.

Foi também implementado um método “summary” para objectos da classe “BIV”. Este método devolve um objecto de classe “summary.BIV” que é uma versão essencialmente resumida de um objecto de classe “BIV”. Eis um exemplo que recorre ao estimador Kaplan-Meier pesado:

```
> KMW <- bivKMW(object=bladderBIV_obj, t1=c(5, 10), t2=c(10, 20), conf=TRUE, conf.level=0.9)

> summary(KMW)

Kaplan-Meier Weighted bivariate probabilities
```

Estimate of P(time1 <= t1, time2 <= t2)

	t2 = 10	t2 = 20
t1 = 5	0.07614291	0.1921058
t1 = 10	0.14385567	0.2598186

Bootstrap confidence band with 1000 samples

5%

	t2 = 10	t2 = 20
t1 = 5	0.02597403	0.1189920
t1 = 10	0.08063223	0.1804715

95%

	t2 = 10	t2 = 20
t1 = 5	0.1313170	0.2755706
t1 = 10	0.2162286	0.3492799

Foi criada uma função cuja finalidade é a simulação de variáveis pseudo-aleatórias[8] bivariadas. Foram implementados dois modelos de probabilidade bivariados distintos, tendo um deles função de distribuição bivariada Weibull dada pela equação (1.5.16), e o outro função de distribuição bivariada exponencial dada pela equação (1.5.19). Os métodos de simulação de variáveis pseudo-aleatórias multivariadas partem de resultados da teoria de simulação univariada. Por isso aconselha-se o estudo prévio dos métodos de simulação univariados. Os métodos de simulação univariados de particular interesse são o método da função de distribuição inversa e o método de transformação descritos em (Johnson M. E., 1987, pp. 19-24). O método utilizado para a geração do vector de variáveis aleatórias relativo ao modelo de probabilidade exponencial bivariado de Farlie-Gumbel-Morgenstern, foi o método da função de distribuição condicional. Uma descrição deste método pode ser encontrada em (Johnson M. E., 1987, pp. 43-45). Resultados do método da

[8] Como não é possível gerar variáveis verdadeiramente aleatórias, é utilizado um algoritmo determinístico partindo de um número frequentemente associado ao tempo da máquina. Desta forma são geradas variáveis aparentemente aleatórias para o utilizador. Por esta razão estas variáveis são denominadas pseudo-aleatórias.

função de distribuição condicional aplicado a modelos de probabilidade bivariados da família de Farlie-Gumbel-Morgenstern podem ser encontrados em (Johnson M. E., 1987, pp. 180-185). Para o modelo de probabilidade bivariado Weibull foi utilizado o método de transformação descrito em (Johnson M. E., 1987, pp. 45-46). A representação de variáveis aleatórias bivariadas Weibull em termos de variáveis aleatórias independentes, um passo crítico para o método de transformação, encontra-se descrita em (Lu & Bhattacharyya, 1990, pp. 554-555). Além disso, uma descrição mais acessível dos resultados do método de transformação aplicado ao modelo de probabilidade bivariado Weibull anteriormente referido, pode ser consultada em (Johnson, Evans, & Green, 1999, p. 5). O modelo de censura pela direita descrito no início da secção 2 foi aplicado. Assim será de seguida descrito por passos, o algoritmo de geração de números aleatórios bivariados que resulta da aplicação dos métodos referidos:

1) Escolher o modelo de probabilidade bivariado, Weibull ou exponencial;
2) Se Weibull, definir os valores dos parâmetros de entrada $\delta, \beta_1, \theta_1, \beta_2, \theta_2$ nos limites especificados na equação (1.5.13), e ir para o passo 4);
3) Se exponencial, definir os valores dos parâmetros de entrada $\alpha, \theta_1, \theta_2$ nos limites especificados na equação (1.5.19), e ir para o passo 9);
4) Gerar U_1, U_2, U_3, U_4, U_5 independentes uniformes no intervalo $[0,1]$;
5) Se $U_5 > \delta$ então $V = -\ln(U_4)$, caso contrário $V = -\ln(U_2) - \ln(U_3)$;
6) $T_1 = {U_1}^{\delta/\beta_1} V^{1/\beta_1} \theta_1$;
7) $T_2 = (1 - U_1)^{\delta/\beta_2} V^{1/\beta_2} \theta_2$;
8) Ir para o passo 14);
9) Gerar U_1, V independentes uniformes no intervalo $[0,1]$;
10) $A = \alpha(2U_1 - 1)$;
11) $U_2 = \frac{2V}{1-A+\sqrt{(1-A)^2+4AV}}$;
12) $T_1 = -\theta_1 \ln(1 - U_1)$;
13) $T_2 = -\theta_2 \ln(1 - U_2)$;
14) Escolher o modelo de censura aleatória, uniforme ou exponencial;
15) Se uniforme, definir o valor do parâmetro de entrada $\lambda \geq 0$, ir para o passo 17);
16) Se exponencial definir o valor do parâmetro de entrada $\lambda > 0$, ir para o passo 18);
17) Gerar C uniforme no intervalo $[0, \lambda]$, ir para o passo 20);
18) Gerar U uniforme no intervalo $[0,1]$;
19) $C = -\lambda \ln(1 - U)$;
20) $Y_1 = \min(T_1, C)$;

21) $\Delta_1 = I(T_1 \leq C)$;

22) $Y_2 = \Delta_1 \min(T_2, C - T_1)$;

23) $\Delta_2 = I(T_2 \leq C - T_1)$;

O algoritmo descrito assume a disponibilidade de uma função capaz de gerar números pseudo-aleatórios uniformes. Este algoritmo foi implementado na função "dgpBIV". Segue-se um exemplo de utilização para o modelo de probabilidade exponencial bivariado:[9]

```
> dgpBIV(n=5, corr=1, dist="exponential", dist.par=c(1, 1), model.cens="uniform", cens.par=3, to.data.frame=TRUE)
```

	time1	event1	time2	event2
1	0.70578365	1	0.15653539	1
2	0.21887652	1	0.03232979	0
3	0.89020808	0	0.00000000	0
4	0.35286195	1	0.74752050	1
5	0.09144789	1	0.22156563	1

Um exemplo para o modelo de probabilidade Weibull bivariado, onde é devolvido um objecto com duas classes, "survBIV" e "weibull":

```
> dgpBIV(n=5, corr=1, dist="weibull", dist.par=c(2, 7, 2, 7), model.cens="exponential", cens.par=6)
```

$data

	time1	event1	time2	event2	Stime
1	1.4815438	0	0.000000	0	1.481544
2	0.7997215	1	1.432653	1	2.232374
3	2.4663818	1	2.821348	1	5.287730
4	2.7370164	0	0.000000	0	2.737016
5	1.2928053	1	2.338680	0	3.631485

$corr

[1] 1

[9] Caso o leitor tente reproduzir estes resultados, é natural que não obtenha os mesmos valores. Uma vez que se tratam de números pseudo-aleatórios, o leitor teria que ter conhecimento do valor da *seed* utilizado para gerar os números pseudo-aleatórios apresentados. Sabendo o valor da *seed* utlizado, seria possível reproduzir os resultados aqui apresentados recorrendo à função "set.seed" disponível no *software* R. Para mais informação, sugere-se a consulta da página de ajuda da função "set.seed", acessível através do *software* R.

```
$dist.par

[1]  2  7  2  7


attr(,"class")

[1]  "survBIV"  "weibull"
```

Os objectos devolvidos pela função “dgpBIV” podem ser guardados para posterior utilização noutras funções aqui apresentadas. O que é particularmente útil para estudos de simulação como o apresentado neste trabalho.

Para calcular os valores teóricos das funções de distribuição bivariadas Weibull e exponencial foi desenvolvida a função “bivBIV”. Esta função devolve um objecto da classe “BIV” que pode ser guardado e utilizado em métodos gráficos. A título de exemplo esta função foi utilizada na criação dos gráficos da Figura 1.4 e da Figura 1.5. Segue-se um exemplo de utilização para o caso de uma função de distribuição exponencial bivariada:

```
> expBIV <- dgpBIV(n=100, corr=1, dist="exponential", dist.par=c(1, 1), model.cens="uniform", cens.par=3, to.data.frame=FALSE)

> grid <- c(0.2231, 1.6094)

> bivBIV( object=expBIV, t1=grid, t2=grid, lower.tail=c(TRUE, TRUE) )

Bivariate exponential probabilities


P(time1 <= 0.2231, time2 <= 0.2231) = 0.0655793749110017


P(time1 <= 0.2231, time2 <= 1.6094) = 0.185567993110251


P(time1 <= 1.6094, time2 <= 0.2231) = 0.185567993110251


P(time1 <= 1.6094, time2 <= 1.6094) = 0.665589323715971
```

Os valores teóricos das funções de distribuição bivariadas podem também ser obtidos sob a forma matricial. Neste caso pode ser utilizada a função "probBIV". Esta função foi implementada devido à sua utilidade em estudos de simulação. Segue-se um exemplo utilizado no estudo de simulação apresentado neste trabalho:

```
> grid <- c(0.2231, 0.5108, 0.9163, 1.6094, 2.3026, 2.9957)

> probBIV(dist="exponential", corr=1, dist.par=c(1, 1), t1=grid, t2=grid)
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	0.06557937	0.1183775	0.1583747	0.1855680	0.1943669	0.1975658
[2,]	0.11837748	0.2175862	0.2975913	0.3583853	0.3815862	0.3913850
[3,]	0.15837470	0.2975913	0.4176041	0.5183994	0.5616039	0.5814029
[4,]	0.18556799	0.3583853	0.5183994	0.6655893	0.7343946	0.7675920
[5,]	0.19436687	0.3815862	0.5616039	0.7343946	0.8181025	0.8592750
[6,]	0.19756582	0.3913850	0.5814029	0.7675920	0.8592750	0.9047533

O coeficiente de correlação depende dos parâmetros do modelo de probabilidade bivariado. A função "corrBIV" permite obter este coeficiente. Por exemplo para uma função de distribuição bivariada Weibull:

```
> corrBIV( dist="weibull", corr=0.5, dist.par=c(2, 7, 2, 7) )

[1]  0.6600091
```

Em relação à função de distribuição exponencial bivariada, o coeficiente de correlação é independente do argumento "dist.par", dependendo apenas do argumento "corr". No entanto por uma questão de coerência com a função de distribuição Weibull bivariada, o argumento "dist.par" deve ser especificado:

```
> corrBIV( dist="exponential", corr=1, dist.par=c(1, 1) )

[1]  0.25
```

3.2 Utilização gráfica

Cada uma das funções gráficas desenvolvidas constitui um método para objectos da classe "BIV". Podendo cada uma ser chamada pelo seu nome genérico. Assim as funções gráficas implementadas na extensão survivalBIV são por ordem alfabética; "contour.BIV", "image.BIV",

"lines.BIV", "persp.BIV" e "plot.BIV". Todas estas funções requerem objectos da classe "BIV" como argumento. Os objectos da classe "BIV" podem ser obtidos recorrendo às funções; "bivBIV", "bivCKM", "bivKMPW", "bivKMW" e "bivLIN". Que por sua vez requerem objectos da classe "survBIV" como argumento. Objectos da classe "survBIV" podem ser obtidos recorrendo às funções "survBIV" e "dgpBIV". Sendo que a função "survBIV" cria um objecto da classe "survBIV" a partir de um objecto do tipo "data.frame". A função "dgpBIV" devolve um objecto da classe "survBIV" caso o argumento "to.data.frame" for igual a "FALSE", caso contrário devolve um objecto do tipo "data.frame". Em seguida serão apresentados exemplos de utilização das funções gráficas mencionadas, bem como os gráficos produzidos pelas mesmas. Alguns destes exemplos constam no manual da extensão survivalBIV (Moreira, Araújo, & Machado, 2012). Sempre que for oportuno, será realizada uma breve interpretação dos gráficos apresentados. Assim pode ser justificado o desenvolvimento e a utilização das funções gráficas implementadas.

Inicialmente é necessário criar e guardar um objecto da classe "survBIV". A base de dados "bladderBIV" é utilizada como exemplo:

```
> bladderBIV_obj <- with( bladderBIV, survBIV(time1, event1, time2, event2) )
```

O objecto "bladderBIV_obj" é guardado para ser utilizado em alguns dos exemplos seguintes.

A função "plot.BIV" cria gráficos de dispersão das probabilidades em relação ao tempo. O argumento "plot.type" permite controlar qual dos intervalos de tempo, o primeiro (time1) ou o segundo (time2), serve de referência para as probabilidades, sendo a variável representada no eixo das abcissas. Esta função, tal como as apresentadas em seguida, pode ser chamada pelo seu nome completo. No entanto, dado cada uma se tratar de um método, é comum ser chamada pelo seu nome genérico, que neste caso será apenas "plot". O exemplo seguinte cria um gráfico de dispersão legendado das estimativas das probabilidades em relação ao segundo intervalo de tempo, para cada um dos valores do primeiro intervalo de tempo. Sendo as bandas de confiança a 95%:

```
> KMPW <- bivKMPW(object=bladderBIV_obj, t1=c(5, 10), t2=c(20, 40), conf=TRUE)

> plot(KMPW, plot.type="t2", lty=1, col=1:2, conf.int=TRUE, legend=TRUE)
```

Neste exemplo é utilizado o estimador Kaplan-Meier pesado pré-suavizado. Correndo as duas linhas de código, é produzido o gráfico da Figura 3.1.

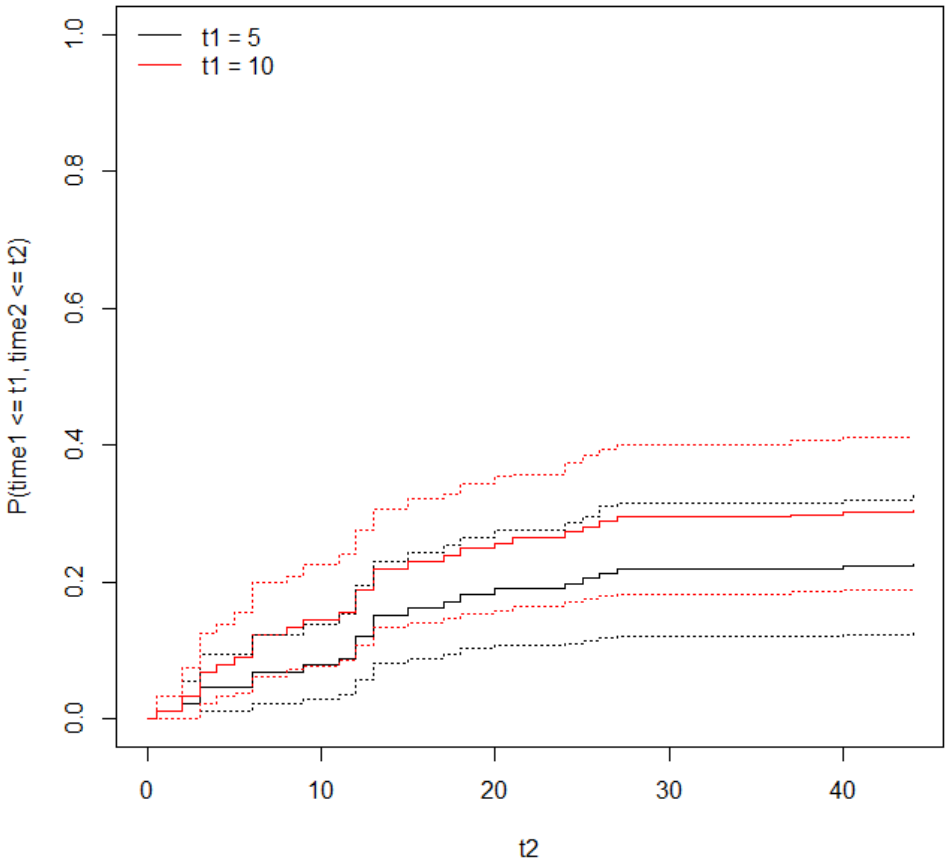


Figura 3.1: Funções de distribuição do segundo intervalo de tempo com bandas de confiança a 95%.

Gráficos deste tipo, para além de proporcionarem uma visão geral dos resultados, podem ter outras utilidades, permitindo tirar algumas conclusões em relação às estimativas representadas. A título de exemplo, o gráfico da Figura 3.1 proporciona um método gráfico para a realização de um teste de hipóteses. Verifica-se que ambas as curvas das estimativas se encontram no interior de ambas as bandas de confiança. Sugerindo que as curvas das estimativas podem ser iguais, a um nível de confiança de 95%. Caso se pretenda a função de distribuição marginal do segundo intervalo de tempo, pode-se especificar o argumento "plot.type" igual a "mt2". Assim utilizando o objecto da classe "BIV" obtido anteriormente:

```
> plot(KMPW, plot.type="mt2", conf.int=TRUE)
```

Correndo a referida função, é obtido o gráfico da Figura 3.2.

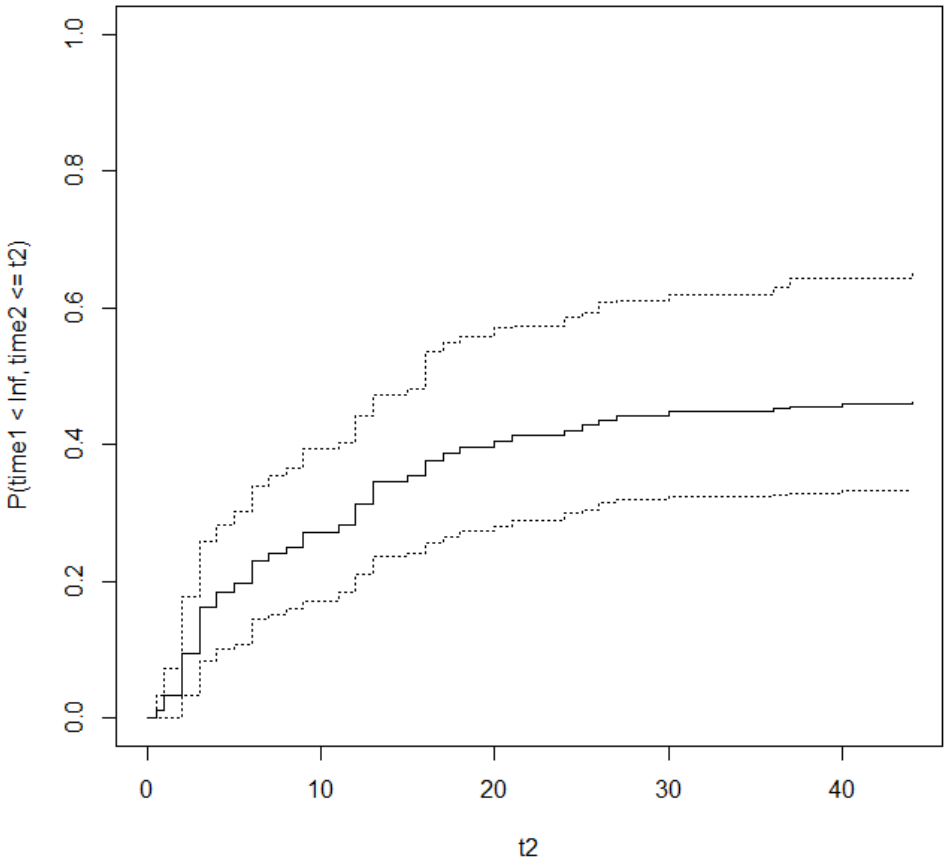


Figura 3.2: Função de distribuição marginal do segundo intervalo de tempo com bandas de confiança a 95%.

O método "lines.BIV" permite adicionar linhas a gráficos existentes. O que é muito útil para exibir curvas de diferentes objectos de classe "BIV" num único gráfico, permitindo a sua comparação. No exemplo seguinte, primeiramente são guardados os objectos "BIV" distintos, cada um relativo a um estimador diferente:

```
> CKM <- bivCKM(object=bladderBIV_obj, conf=FALSE)

> KMPW <- bivKMPW(object=bladderBIV_obj, conf=FALSE)

> KMW <- bivKMW(object=bladderBIV_obj, conf=FALSE)

> LIN <- bivLIN(object=bladderBIV_obj, conf=FALSE)
```

Em seguida é criado o gráfico de dispersão em relação a cada um dos estimadores. Neste caso é criado o gráfico da função de distribuição marginal do segundo intervalo de tempo:

```
> plot(CKM, plot.type="mt2", lty=1, col=1)
```

Finalmente, recorrendo à função “lines”, são adicionadas as curvas das funções de distribuição marginais do segundo intervalo de tempo, relativas a cada um dos estimadores restantes:

```
> lines(KMPW, plot.type="mt2", lty=1, col=2)

> lines(KMW, plot.type="mt2", lty=1, col=3)

> lines(LIN, plot.type="mt2", lty=1, col=4)
```

A adição de uma legenda permite associar facilmente cada curva ao seu estimador. O resultado final é o gráfico da Figura 3.3.

```
> legend(x="topleft", legend=c("CKM", "KMPW", "KMW", "LIN"), lty=1, col=1:4, bty="n")
```

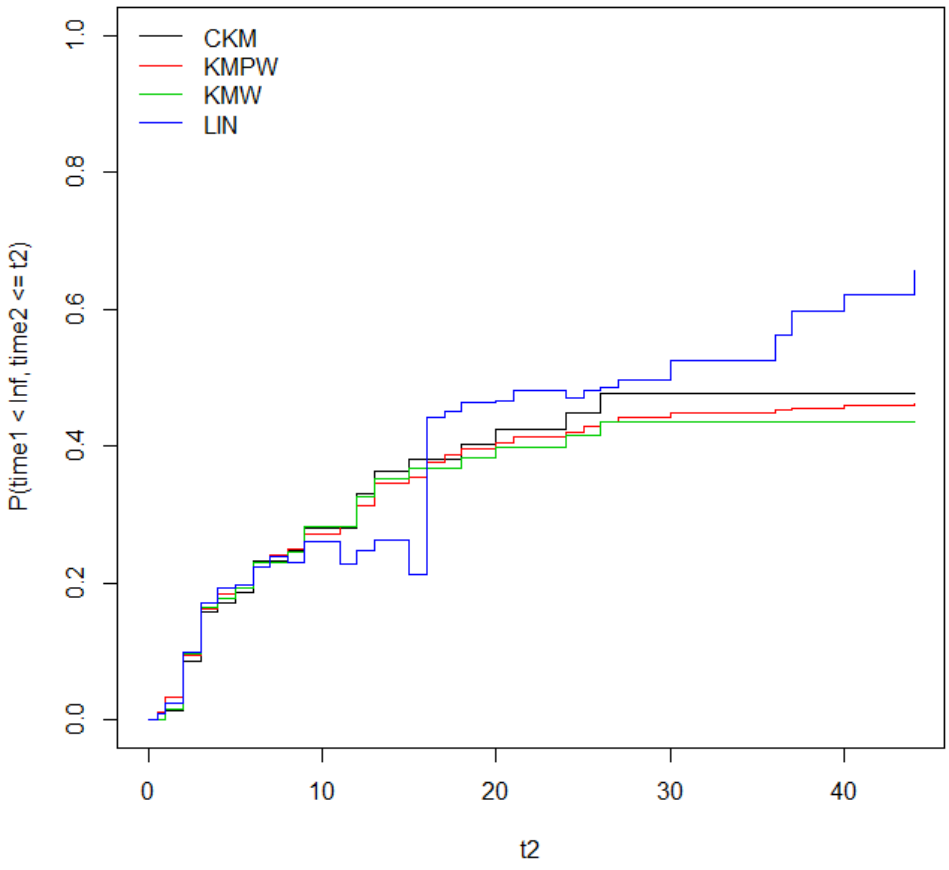


Figura 3.3: Funções de distribuição marginais do segundo intervalo de tempo.

O gráfico da Figura 3.3 permite comparar as estimativas obtidas por cada um dos estimadores. Neste caso verifica-se que a curva de distribuição marginal do segundo intervalo de tempo obtida pelo estimador de Lin se destaca das restantes curvas. Sendo que as curvas relativas aos estimadores Kaplan-Meier pesado e Kaplan-Meier pesado pré-suavizado se encontram muito próximas. Este é apenas um dos muitos exemplos das constatações que se pode fazer recorrendo a gráficos deste tipo.

A função "contour.BIV" cria gráficos de linhas de nível das probabilidades. No exemplo seguinte é simulada uma amostra de 500 observações, sendo guardada no objecto da classe "survBIV" de nome "BIVdata". A seguir é criado um objecto da classe "BIV" chamado "KMPW", onde constam as estimativas das probabilidades resultantes da aplicação do estimador Kaplan-Meier pesado pré-suavizado. O objecto "EXP" da classe "BIV" guarda as probabilidades teóricas:

```
> BIVdata <- dgpBIV(n=500, corr=1, dist="exponential", dist.par=c(1, 1), model.cens="uniform", cens.par=4)

> KMPW <- bivCKM(object=BIVdata, conf=FALSE)

> EXP <- bivBIV(object=BIVdata)
```

Recorrendo aos objectos da classe "BIV" guardados anteriormente, é possível sobrepor as linhas de nível relativas a cada um dos objectos:

```
> contour(KMPW, col=1, lty=1, ylim=c( min(KMPW$gridy), max(KMPW$gridy)+0.1 ), main="P(time1 <= t1, time2 <= t2)", legend=FALSE)

> contour(EXP, col=2, lty=1, add=TRUE, legend=FALSE)
```

A adição de uma legenda permite identificar as linhas de nível:

```
> legend("topleft", legend=c("KMPW", "exponential"), col=1:2, lty=1, bty="n", ncol=2)
```

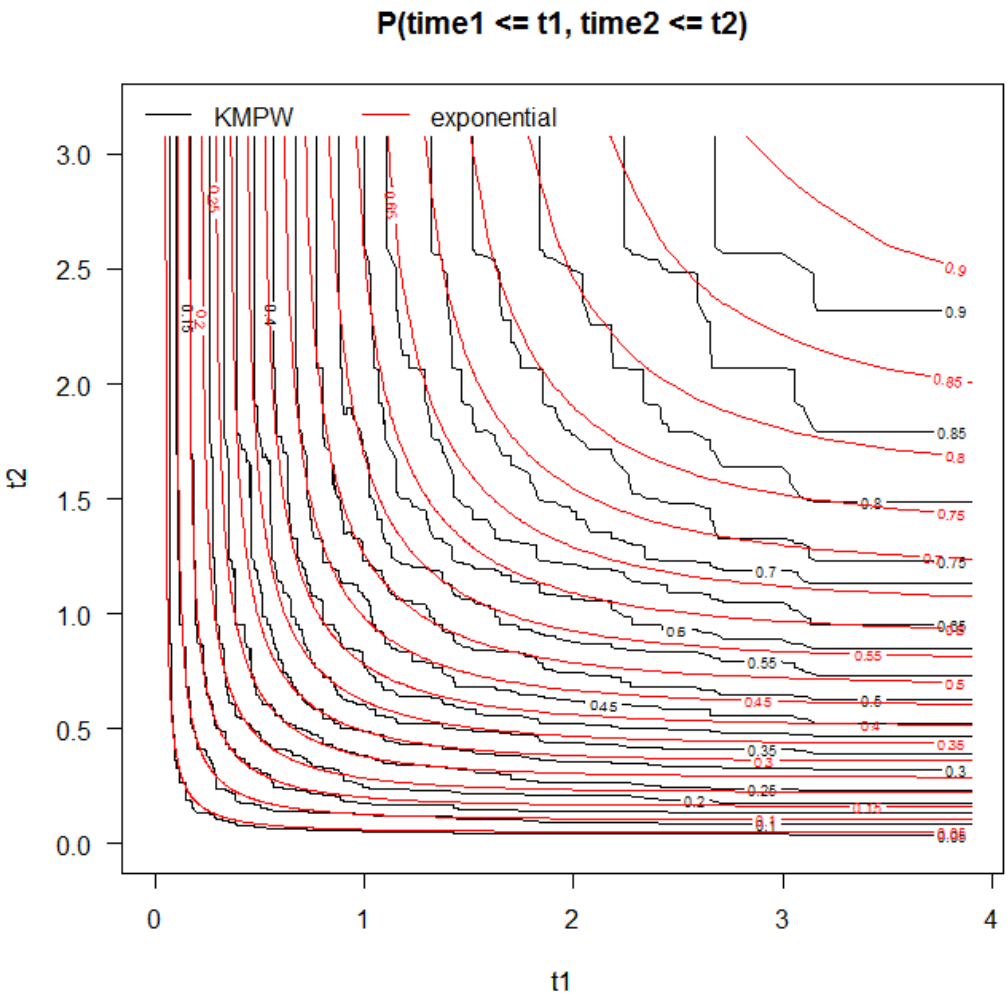


Figura 3.4: Linhas de nível das probabilidades estimadas (KMPW) e teóricas (exponential).

Correndo as 6 linhas de código mencionadas neste parágrafo, obtém-se o gráfico da Figura 3.4. De modo a interpretar o gráfico da Figura 3.4 é necessário definir previamente linhas de nível. Tomando como exemplo o gráfico da Figura 3.4, linhas de nível são linhas que unem pontos de coordenadas $(t1, t2)$ para os quais a probabilidade é igual. No mesmo gráfico, o valor da probabilidade encontra-se adjacente à respectiva linha de nível. Assim o gráfico da Figura 3.4 permite observar que à medida que aumentam os valores das coordenadas $t1$ e $t2$, a estimativa afasta-se do valor teórico. Para que a estimativa seja considerada uma boa estimativa, deve estar o mais próxima possível do verdadeiro valor.

O método "image.BIV" produz um gráfico idêntico ao criado pelo método "contour.BIV". Neste caso as distintas regiões limitadas por linhas de nível são preenchidas com diferentes cores, o

que facilita a observação e interpretação do gráfico. No exemplo seguinte são criadas imagens com linhas de nível para as estimativas bem como para as regiões de confiança:

<- bivLIN(object=bladderBIV_obj, conf=TRUE)

> image(LIN, conf.int=TRUE, text.col="blue")

O gráfico da Figura 3.5 sai como resultado. Este gráfico permite uma visualização relativamente global das probabilidades, condensando uma grande quantidade de informação.

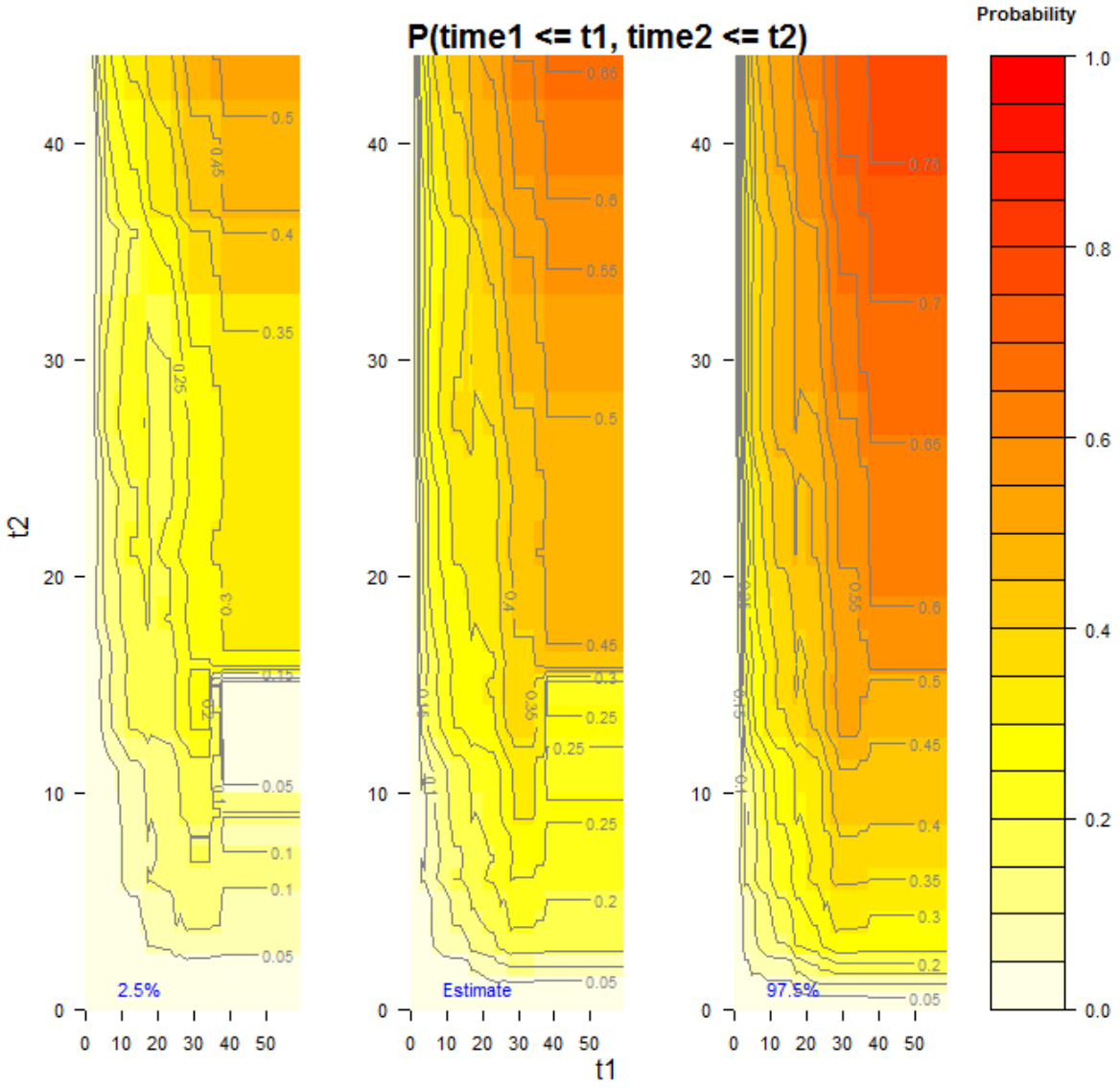


Figura 3.5: Imagens das probabilidades estimadas com regiões de confiança 95%.

A função “persp” permite representar dados tridimensionais em perspectiva. Esta função desenha uma superfície num referencial cartesiano tridimensional. O exemplo seguinte cria a perspectiva da função de distribuição bivariada estimada no centro, com a perspectiva da região de confiança inferior do lado esquerdo, e a perspectiva da região de confiança superior do lado direito:

```
> dev.off()

null device
          1
> windows(record=TRUE, width=20*1.6, height=20)

> KMW <- bivKMW(object=bladderBIV_obj, conf=TRUE)

> persp(KMW, persp.type="t2", conf.int=TRUE, cex=1.5, cex.lab=1)
```

Neste caso a função “windows” serve para redimensionar a janela gráfica, sendo produzida uma imagem de resolução superior à resolução padrão. Os gráficos da Figura 3.6 são criados.

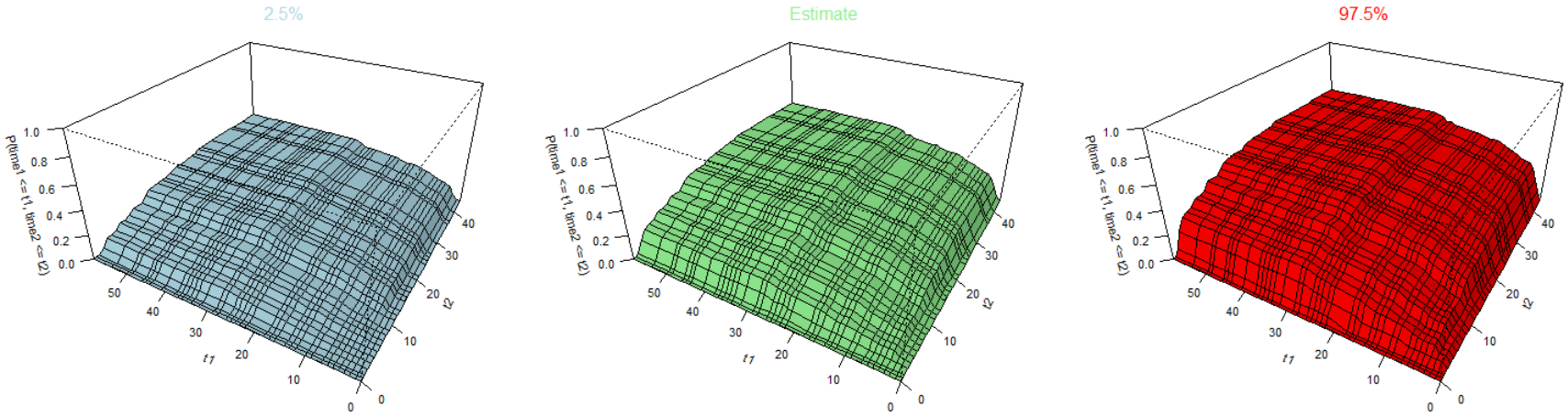


Figura 3.6: Perspectivas da função de distribuição bivariada estimada e das regiões de confiança a 95%.

Mais um exemplo de utilização do método “persp.BIV” será apresentado. Neste caso é simulada uma amostra de 100 observações que fica guardada no objecto da classe “survBIV” nomeado “BIVdata”. A partir deste objecto são criados os objectos da classe “BIV” da estimativa de Lin chamado “LIN”, e dos valores teóricos das probabilidades chamado “EXP”:

```
> BIVdata <- dgpBIV(n=100, corr=1, dist="exponential", dist.par=c(1, 1), model.cens="uniform", cens.par=4)

> LIN <- bivLIN(object=BIVdata, conf=FALSE)

> EXP <- bivBIV(object=BIVdata)
```

As perspectivas da função de distribuição bivariada teórica e estimada são desenhadas num único dispositivo gráfico recorrendo aos comandos seguintes:

```
> dev.off()

null device
          1
> windows(record=TRUE, width=20*1.6, height=20)

> par( mfrow=c(1, 2) )

> persp(EXP, legend=TRUE, text.col="blue", inset=0.1)

> persp(LIN, legend=TRUE, text.col="blue", inset=0.1)
```

Mais uma vez é utilizada a função “windows” de modo a aumentar a resolução da imagem pretendida. A execução das 8 linhas de código anteriores cria os gráficos da Figura 3.7. Nesta figura a superfície suavizada da função de distribuição bivariada teórica contrasta com a superfície irregular da função de distribuição bivariada estimada.

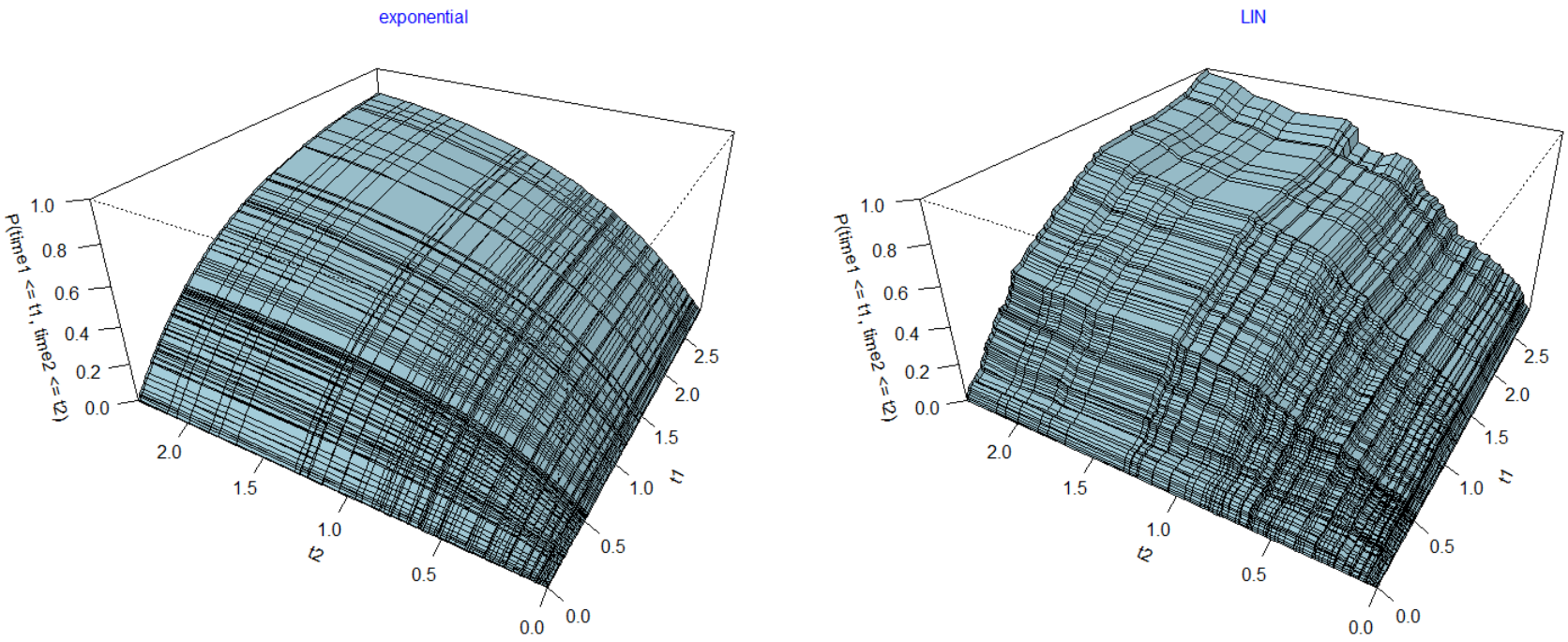


Figura 3.7: Perspectivas das funções de distribuição bivariada exponencial e estimada pelo estimador de Lin.

Esta página foi intencionalmente deixada em branco.

4 Estudo de simulação

Recorrendo aos métodos numéricos e gráficos introduzidos na secção anterior, bem como aos fundamentos teóricos resumidos na secção 1, serão comparados os estimadores abordados na secção 2. O objectivo é seleccionar o estimador que apresenta o melhor comportamento em termos das propriedades desejáveis dos estimadores. É desejável que os estimadores proporcionem resultados próximos dos verdadeiros valores, ou seja que apresentem erros quadráticos médios próximos de zero. Neste caso, estando perante estimadores de funções, é desejável que os estimadores que são também funções, apresentem as mesmas propriedades das funções que pretendem estimar. Assim inicialmente serão analisados os 4 estimadores da função de distribuição bivariada apresentados na secção 2, em relação às propriedades das funções de distribuição bivariadas referidas em (1.5.2). De modo a ilustrar a violação de uma qualquer propriedade entre as cinco propriedades referidas, serão apresentados gráficos de probabilidades estimadas a partir de amostras simuladas, sempre que for oportuno. Em seguida os mesmos estimadores serão analisados em relação às propriedades desejáveis dos estimadores, referidas na secção 1.3. Para o efeito foram simuladas 10000 amostras distintas e independentes para cada um de dois cenários de censura, e para cada um de dois cenários de correlação. Para cada amostra foram obtidas as estimativas da função de distribuição bivariada em relação a vários quantis, e para cada um dos estimadores apresentados na secção 2. O viés, desvio padrão e o erro quadrático médio foram calculados e registados, tendo sido construídas as tabelas e as figuras em anexo.

4.1 Propriedades da função de distribuição bivariada

4.1.1 Estimador Kaplan-Meier condicional

O estimador Kaplan-Meier condicional apresentado em (2.1.1) é uma função monótona não decrescente na componente y. Pois quando se estabelece um valor constante da variável x, o mesmo torna-se no estimador Kaplan-Meier da função de distribuição univariada do segundo intervalo de tempo, apresentando o mesmo comportamento. O mesmo pode não se verificar em relação à componente x. Sendo estabelecido um valor constante para a variável y, a cada vez que é avaliado o estimador para cada valor diferente da variável x, ocorre uma redistribuição da amostra de forma a condicionar a mesma ao conjunto $Y_{1i} \leq x$, podendo resultar um decréscimo da função

estimada. De forma a ilustrar esta eventualidade foi simulada uma amostra de 50 observações, a partir da qual foi construído o gráfico da Figura 4.1.

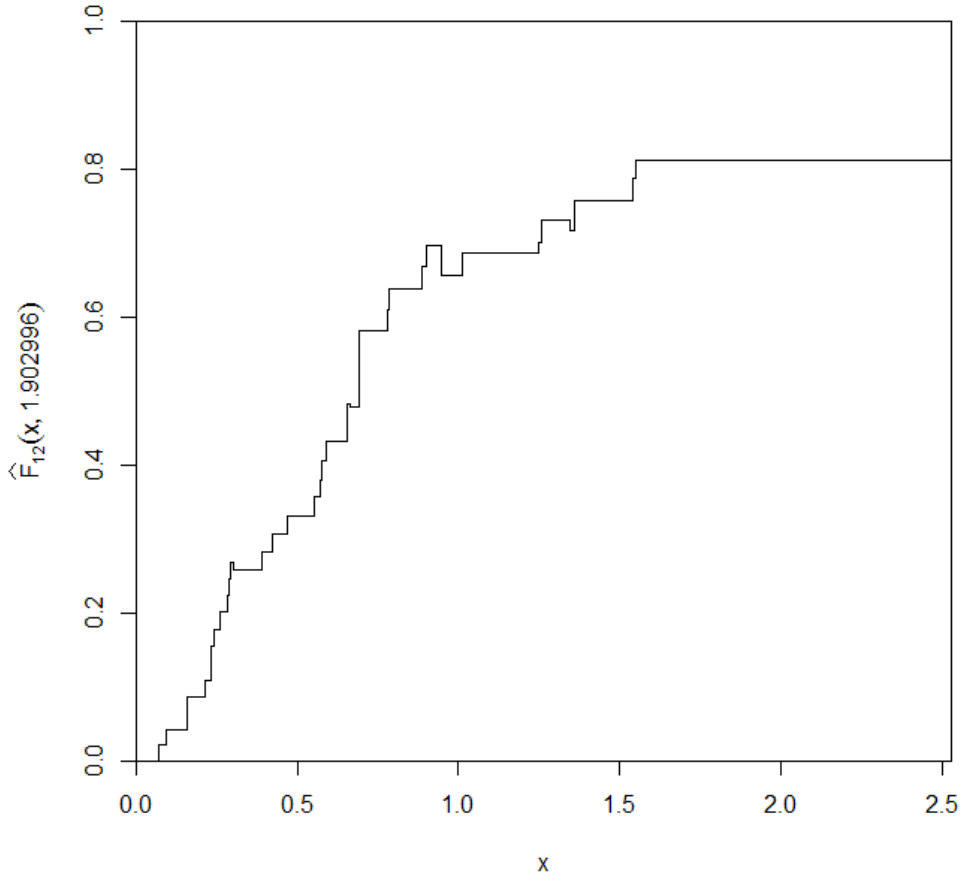


Figura 4.1: Função de distribuição marginal obtida à custa do estimador Kaplan-Meier condicional.

No gráfico da Figura 4.1 observa-se um decréscimo da função de distribuição univariada em pelo menos 4 valores distintos da variável representada no eixo das abcissas. Deve-se assinalar que esta situação depende da amostra, podendo não ocorrer. Neste caso foi necessário repetir a simulação várias vezes, de modo a obter a amostra que deu origem ao gráfico apresentado na Figura 4.1. Pode-se assim constatar que o estimador Kaplan-Meier condicional da função de distribuição bivariada não garante a propriedade (1) referida em (1.5.2).

O estimador Kaplan-Meier da função de sobrevivência não pode ser negativo ou superior à unidade. Pelo que facilmente se verifica que o estimador Kaplan-Meier condicional da função de

distribuição bivariada dado pela expressão (2.1.1), também não pode ser negativo ou exceder a unidade, garantindo assim a propriedade (2) mencionada em (1.5.2).

Em relação ao estimador Kaplan-Meier da função de sobrevivência $\hat{S}(.)$ é sabido que $\hat{S}(-\infty) = 1$, em consequência da expressão (1.4.2). Fazendo a substituição em (2.1.1) vem:

$$\hat{F}_{12}(-\infty, y) = \left[1 - \hat{S}_1(-\infty)\right]\left[1 - \hat{S}_2(y|T_1 \leq -\infty)\right] = [1-1]\left[1 - \hat{S}_2(y|T_1 \leq -\infty)\right] = 0$$

$$\hat{F}_{12}(x, -\infty) = \left[1 - \hat{S}_1(x)\right]\left[1 - \hat{S}_2(-\infty|T_1 \leq x)\right] = \left[1 - \hat{S}_1(x)\right][1-1] = 0$$

Pode-se assim concluir que o estimador Kaplan-Meier condicional exibe a propriedade (3) enumerada no conjunto de expressões (1.5.2).

É sabido que o estimador Kaplan-Meier da função de sobrevivência pode não atingir o valor nulo. Caso todas as observações associadas ao tempo máximo registado na amostra, forem todas censuradas, o estimador Kaplan-Meier proporciona um resultado maior que zero. Nesta situação pode ocorrer $\hat{S}_1(+\infty) \neq 0$ ou $\hat{S}_2(+\infty|T_1 \leq +\infty) \neq 0$, o que implicaria $\hat{F}_{12}(+\infty, +\infty) \neq 1$. Sendo assim inviabilizada a propriedade (4) enumerada em (1.5.2).

Relativamente à propriedade (5) do conjunto de propriedades (1.5.2), a verificação da mesma pode apresentar dificuldades em termos da equação apresentada para a probabilidade do rectângulo. Felizmente pode ser definida uma equação equivalente, que por sua vez facilita a validação da mesma propriedade. No seguimento da equação (1.5.5), a probabilidade do rectângulo pode ser obtida recorrendo à seguinte equação:

$$P[rectângulo] = \sum_{x_1 \leq x_i \leq x_2} \sum_{y_1 \leq y_i \leq y_2} f(x_i, y_i) \qquad (4.1.1)$$

onde $f(x_i, y_i)$ representa a função massa de probabilidade avaliada nos pontos de coordenadas (x_i, y_i). Graficamente a função massa de probabilidade é igual ao salto observado no ponto ao qual a mesma se refere. O gráfico da Figura 4.1 sugere a possibilidade de ocorrência de saltos de valor negativo. Sendo os saltos negativos observados nos pontos em que a curva da função de distribuição decresce. Havendo a possibilidade de ocorrência de valores de $f(x_i, y_i)$ negativos, pode em teoria ser definido um rectângulo para o qual a probabilidade resulta negativa. Não sendo garantida a propriedade (5) do conjunto de propriedades (1.5.2).

4.1.2 Estimador Kaplan-Meier pesado

Relativamente ao estimador Kaplan-Meier pesado definido em (2.2.1), pode verificar-se que o mesmo é uma função monótona não decrescente nas compenentes x e y. Uma vez que $W_i \geq 0$, à medida que os valores das variáveis x e ou y aumentam, $\hat{F}_{12}(x, y)$ não pode decrescer. Pela mesma razão, $\hat{F}_{12}(x, y)$ não pode ser inferior a zero. A soma de todos os pesos Kaplan-Meier W_i nunca pode exceder a unidade. No limite quando x ou y se aproximam de menos infinito, nenhum dos pesos contribui para a estimativa, pelo que a função se aproxima de zero. Devido aos pesos serem todos não negativos, não é possível definir um rectângulo para o qual a probabilidade é negativa. Assim se conclui que o estimador (2.2.1) garante as propriedades (1), (2), (3) e (5) apresentadas em (1.5.2).

Caso todas as observações associadas ao tempo máximo registado na amostra, forem todas censuradas, todos os pesos W_i a elas associados são nulos. Nesta situação a soma de todos os pesos é inferior à unidade. Constata-se deste modo que o estimador Kaplan-Meier pesado não garante a propriedade (4) definida em (1.5.2).

4.1.3 Estimador Kaplan-Meier pesado pré-suavizado

O estimador Kaplan-Meier pesado pré-suavizado garante as mesmas propriedades que a versão não suavizada. Ambos os estimadores são idênticos, pelo que seguindo um raciocínio idêntico ao abordado na subsecção anterior, é possível tirar exactamente as mesmas conclusões.

Uma vez que se pretende comparar todos os estimadores em estudo, para que esta subsecção não seja demasiado curta, serão apontadas diferenças entre os estimadores Kaplan-Meier pesados na sua versão suavizada e não suavizada. Pela análise das expressões para os pesos Kaplan-Meier, utilizadas por cada um dos estimadores, pode verificar-se que a versão suavizada pode pesar positivamente observações para as quais o segundo intervalo de tempo é censurado. Enquanto o estimador Kaplan-Meier pesado atribui peso nulo ao mesmo tipo de observações. Pelo que graficamente é de esperar que a curva da distribuição estimada à custa do estimador Kaplan-Meier pesado pré-suavizado, apresente maior número de saltos.

Uma vez que os valores Δ_{2j} podem assumir apenas valor 0 ou 1, enquanto os valores $m(\tilde{T}_{1j}, \tilde{T}_{2j})$ podem assumir qualquer valor entre 0 e 1, é de esperar menor variância por parte dos

segundos relativamente aos primeiros. Pelo que em princípio o estimador Kaplan-Meier pesado pré-suavizado deverá exibir menor variância que o estimador Kaplan-Meier pesado.

De modo a verificar graficamente as hipóteses anteriormente levantadas, foi simulada uma amostra com 100 observações, a partir da qual foi criado o gráfico da Figura 4.2.

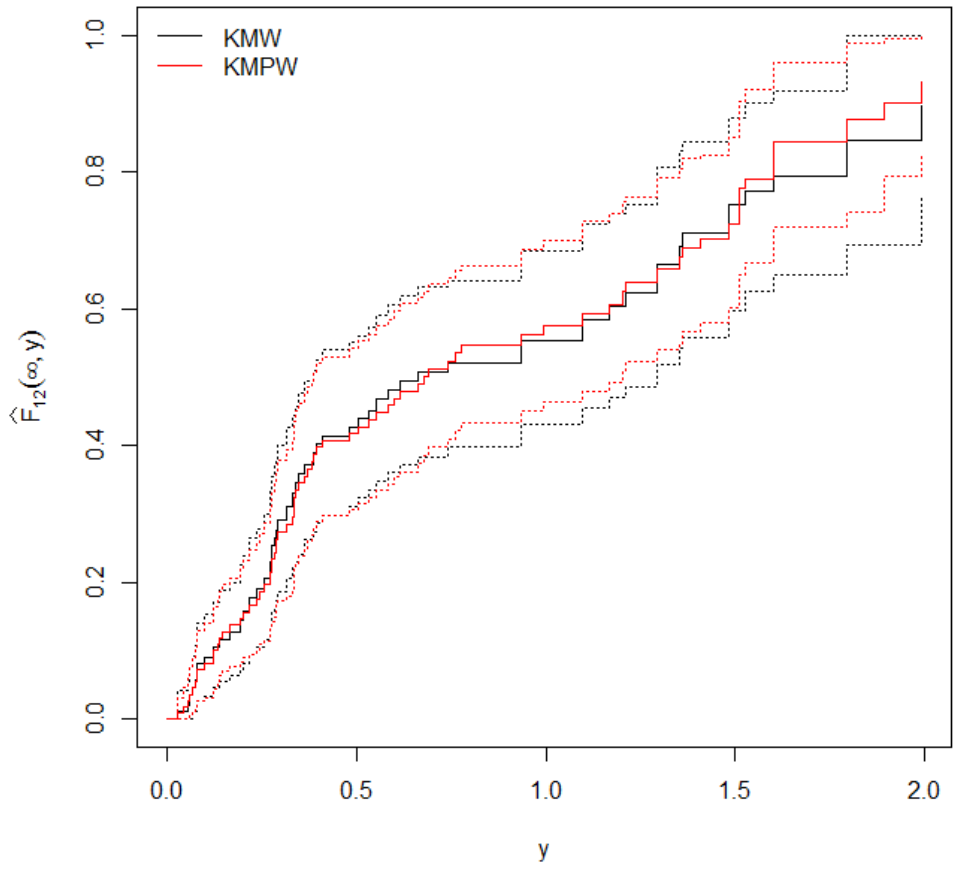


Figura 4.2: Curvas da função de distribuição marginal do segundo intervalo de tempo com bandas de confiança 95%.

Observa-se um maior número de saltos em relação à curva obtida pelo estimador Kaplan-Meier pesado pré-suavizado (KMPW). Além disso as bandas de confiança a 95% relativas à curva obtida pelo estimador Kaplan-Meier pesado pré-suavizado (KMPW) são em geral mais estreitas, o que indica menor variância por parte deste estimador em relação ao estimador Kaplan-Meier pesado (KMW).

4.1.4 Estimador de Lin

O estimador de Lin definido em (2.4.3) não garante $\hat{F}_{12}(x,y)$ não decrescente, nem na componente x nem na componente y. Este facto pode ser melhor ilustrado com uma imagem com linhas de nível. Como tal foi simulada uma amostra com 200 observações, a partir da qual foi criado o gráfico da Figura 4.3.

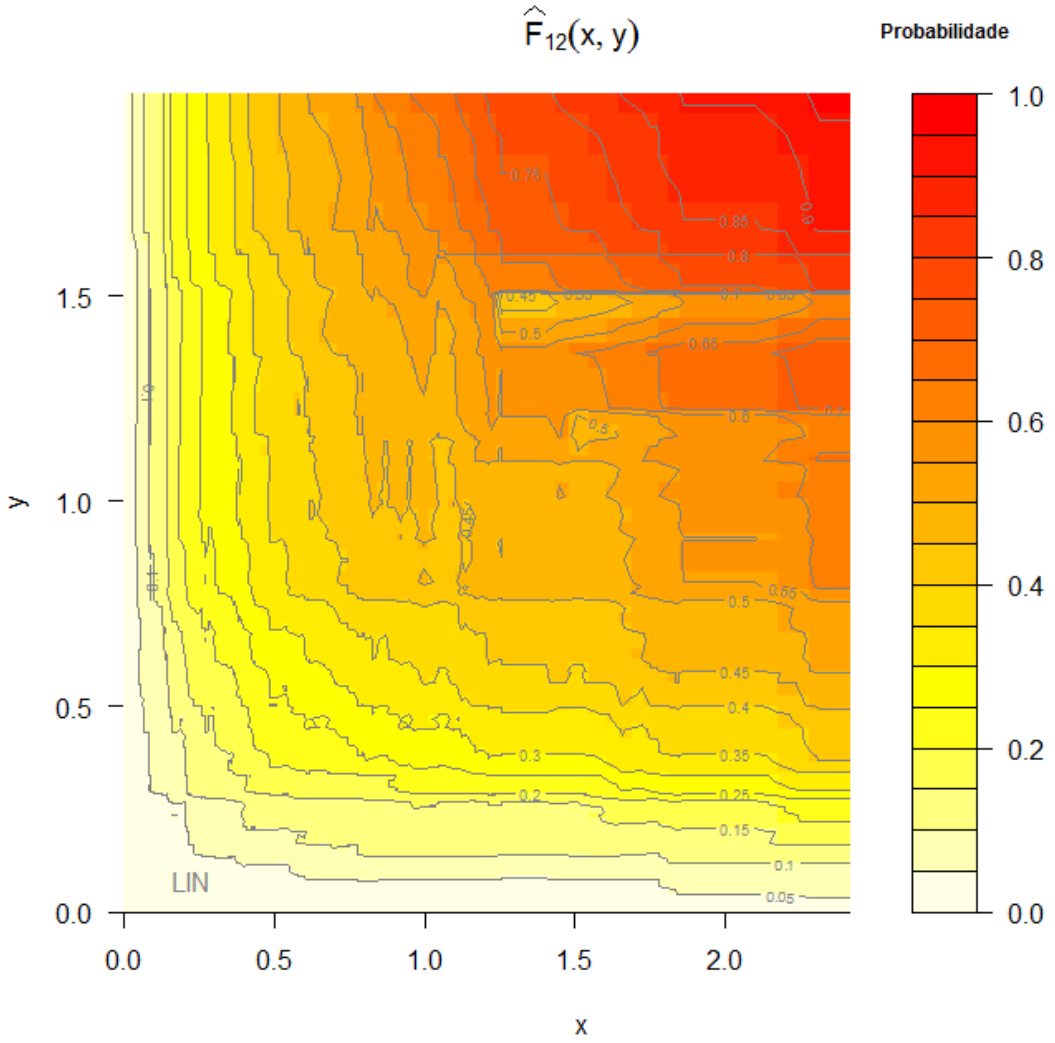


Figura 4.3: Imagem da função de distribuição bivariada obtida pelo estimador de Lin.

Observam-se regiões de baixa intensidade de cor limitadas por regiões envolventes de maior intensidade de cor. O que indica que a função de distribuição bivariada diminui com o aumento de ambas as variáveis dependentes. Havendo uma violação da propriedade (1) definida em (1.5.2).

A propriedade (2) apresentada em (1.5.2) também não pode ser assegurada pelo estimador (2.4.3). Probabilidades negativas podem ocorrer, pois o segundo termo da diferença pode ser maior que o primeiro.

Quando qualquer uma das duas variáveis se aproxima de menos infinito, nenhum dos termos contribui para a soma em (2.4.3). Pelo que a propriedade (3) de (1.5.2) se encontra assegurada, independentemente da amostra.

Substituindo x e y por infinito em (2.4.3) vem $\hat{F}_{12}(+\infty,+\infty) = \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_{1i}}{\hat{G}(Y_{1i})}$, uma expressão idêntica ao estimador Kaplan-Meier na forma (1.4.20). Tal como acontece com o estimador Kaplan-Meier, a soma será inferior à unidade quando as observações associadas ao maior tempo na amostra forem todas censuradas. Não havendo garantias em relação à propriedade (4) definida em (1.5.2).

O gráfico da Figura 4.3 sugere a possibilidade de ocorrência de saltos negativos. Pelo que segundo a equação (4.1.1) a propriedade (5) definida em (1.5.2) também não se encontra assegurada. Uma vez que havendo a possibilidade de ocorrência de saltos negativos, pode ser definido um rectângulo para o qual a probabilidade resulta negativa.

4.2 Enviesamento, desvio padrão e erro quadrático médio

De modo a estudar o comportamento de cada um dos quatro estimadores apresentados na secção 2; em relação ao enviesamento, desvio padrão e erro quadrático médio; foi realizado um estudo de simulação. Foi simulado um determinado número de amostras $(Y_1, Y_2, \Delta_1, \Delta_2)$ independentes entre si, recorrendo ao modelo de probabilidade bivariado de Farlie-Gumbel-Morgenstern com marginais exponenciais padrão ($\theta_1 = 1, \theta_2 = 1$), cuja função de distribuição se encontra definida em (1.5.19). Foram simulados dois cenários de correlação entre as duas variáveis aleatórias, independência ($\alpha = 0$) e correlação máxima ($\alpha = 1$). Sendo a correlação máxima igual a 0.25. Os intervalos de tempo (T_1, T_2) simulados foram sujeitos a dois cenários de censura distintos. A variável aleatória de censura (C) foi simulada de acordo com o modelo de probabilidade uniforme no intervalo $[0,3]$ num caso, e uniforme no intervalo $[0,4]$ no outro. Sendo que o primeiro cenário

de censura resulta em amostras com maior percentagem de observações censuradas.[10] Em relação ao tamanho da amostra foram considerados cinco cenários: 50, 100, 200, 400 e 800. A combinação dos dois cenários para o modelo de probabilidade bivariado exponencial, com os dois cenários para a variável aleatória de censura, com cada um dos tamanhos da amostra referidos, totaliza vinte cenários de amostra distintos. Para cada um dos vinte cenários foram simuladas 10000 amostras independentes, resultando num total de 200000 amostras distintas. Para cada uma das amostras foi estimado o valor da função de distribuição bivariada recorrendo a cada um dos quatro estimadores apresentados na secção 2. Foram utilizados os valores 0.2231, 0.5108, 0.9163, 1.6094, 2.3026 e 2.9957; para cada uma das variáveis x e y. O que resulta em 36 quantis bivariados possíveis. Uma vez que o verdadeiro valor da função de distribuição bivariada pode ser determinado segundo a equação (1.5.19), foi determinado o viés, o desvio padrão e o erro quadrático médio em relação a cada estimador, para cada quantil bivariado e para cada um dos vinte cenários anteriormente mencionados. Estes resultados são apresentados em anexo sob a forma de tabelas e gráficos. Tendo a maioria dos gráficos em anexo sido construídos recorrendo à extensão lattice (Sarkar, 2008) para o *software* R. Por uma questão de espaço as tabelas apresentam apenas parte dos resultados, no entanto os gráficos contemplam todo o conjunto de resultados obtido. O anexo foi dividido por cada um dos quatro estimadores. Assim, o Anexo A contém tabelas e gráficos relativos ao estimador Kaplan-Meier condicional (CKM); o Anexo B agrega os resultados do estimador Kaplan-Meier pesado (KMW); o Anexo C refere-se aos resultados obtidos pelo estimador Kaplan-Meier pesado pré-suavizado (KMPW); referindo-se o Anexo D ao estimador de Lin (LIN). De forma a comparar entre si os quatro estimadores em relação ao erro quadrático médio, foram determinadas as eficiências relativas, tendo sido construídos os gráficos do Anexo E. Os resultados apresentados em anexo serão de seguida alvo de interpretação. Dado que os gráficos contemplam a mesma informação contida nas tabelas, será dada preferência aos gráficos, servindo as tabelas apenas de referência. Além disso, os gráficos facilitam imenso as interpretações, em relação às tabelas. Primeiramente cada um dos quatro estimadores será interpretado em relação ao viés, desvio padrão e erro quadrático médio. Para finalizar será analisada a eficiência relativa, na tentativa de identificar o estimador mais eficiente. Uma vez que haverá a necessidade de referir muitas vezes os estimadores, serão abreviadamente referidos pelos respectivos acrónimos.

O viés de cada estimador foi obtido segundo a equação (1.3.1), pelo que um valor negativo indica que o estimador em estudo está a subestimar o valor verdadeiro de $F_{12}(x, y)$, um viés positivo indica que o mesmo estimador está a sobrestimar o verdadeiro valor de $F_{12}(x, y)$. Portanto

[10] $C \sim U[0,3]$ resulta em amostras com cerca de 32% de observações censuradas em relação ao primeiro intervalo de tempo, e cerca de 56% das observações do segundo intervalo de tempo censuradas. $C \sim U[0,4]$ por sua vez resulta em cerca de 25% das observações do primeiro intervalo de tempo censuradas, e aproximadamente 46% de observações censuradas em relação ao segundo intervalo de tempo.

é desejável que o viés de um estimador seja muito próximo de zero. O desvio padrão é a raiz quadrada da variância definida pela expressão (1.3.6), podendo assumir qualquer valor maior ou igual a zero. Baixos valores de desvio padrão de um estimador indicam que o mesmo se distribui muito próximo do verdadeiro valor de $F_{12}(x, y)$, sendo desejáveis. O erro quadrático médio de cada estimador foi calculado segundo a equação (1.3.7), reflectindo a contribuição do viés e da variância. O erro quadrático médio é não negativo, sendo desejável que se encontre o mais próximo de zero quanto possível.

4.2.1 Estimador Kaplan-Meier condicional

O gráfico da Figura A.1 sugere que quando os intervalos de tempo são independentes, o estimador CKM em média subestima o verdadeiro valor da função de distribuição bivariada $F_{12}(x, y)$. Neste caso foram registados valores de viés mais próximos de zero para menores valores dos quantis x e y. Observa-se uma diminuição do viés à medida que os valores quer de uma variável quer de outra aumentam. No entanto aumentos da variável x registaram maiores decréscimos de viés que aumentos da variável y. Esta observação encontra-se em sintonia com o facto de o estimador CKM não ser uma função de distribuição própria na componente x, sendo-o no entanto na componente y. Sugere-se a revisão da subsecção 4.1.1, onde foram analisadas as propriedades do estimador CKM.

De acordo com o gráfico da Figura A.2, quando os intervalos de tempo apresentam correlação máxima, o estimador CKM em média sobrestima o valor de $F_{12}(x, y)$ na maior parte da região definida pelo par de variáveis (x, y). Sendo que neste cenário de correlação, sob maior percentagem de censura, para valores superiores da variável x, este estimador subestima em média o valor de $F_{12}(x, y)$. Em ambos os cenários de censura, os valores maiores de viés foram registados na região de menores valores de x e maiores valores de y. Tendo a região de maiores valores de ambas as variáveis x e y, registado os menores valores de viés.

Foram registados valores de viés mais próximos de zero no cenário de correlação nula. Em ambos os cenários de correlação foi registada uma aproximação a zero do viés com a diminuição da percentagem de censura e com o aumento da dimensão da amostra. A aproximação a zero do viés com o aumento da dimensão da amostra sugere a verificação da condição de consistência (1.3.10). No entanto trata-se de uma condição insuficiente de consistência, na medida em que existem condições adicionais que se devem verificar em simultâneo para que o estimador possa ser

considerado consistente. Neste caso existe apenas uma condição adicional, que será de seguida analisada.

O gráfico da Figura A.3 apresenta um comportamento em tudo semelhante ao gráfico da Figura A.4. Assim em ambos os cenários de correlação e censura, o desvio padrão de $\hat{F}_{12}(x,y)$ aumenta com o aumento de ambas as variáveis x e y. Registou-se uma aproximação a zero por parte do desvio padrão com a diminuição da percentagem de censura. Observa-se uma tendência para zero por parte do desvio padrão, à medida que aumenta o tamanho da amostra, o que sugere a verificação da condição de consistência (1.3.11). Uma vez que a condição de consistência (1.3.10) é sugerida pelos gráficos da Figura A.1 e da Figura A.2, os quatro gráficos mencionados sugerem que o estimador CKM é consistente, pois as duas condições necessárias de consistência são sugeridas em simultâneo. Os gráficos não evidenciam uma diferenciação significativa do desvio padrão em relação a cada um dos cenários de correlação.

O gráfico da Figura A.5 sugere que no cenário de correlação nula, o erro quadrático médio de $\hat{F}_{12}(x,y)$ segundo o estimador CKM tem tendência a aumentar com o aumento de ambas as variáveis x e y. Tal como o viés de $\hat{F}_{12}(x,y)$ segundo o estimador CKM, também o erro quadrático médio registou um padrão de linhas de nível distinto em relação a cada cenário de correlação. Em relação ao erro quadrático médio de $\hat{F}_{12}(x,y)$ segundo o estimador CKM representado no gráfico da Figura A.6, que se refere ao cenário de correlação máxima, registou-se uma tendência idêntica à registada no gráfico da Figura A.2, que por sua vez se refere ao viés no mesmo cenário de correlação. Assim em ambos os cenários de censura, os maiores valores de erro quadrático médio foram observados na região de menores valores da variável x e maiores valores da variável y. O cenário de maior percentagem de censura registou também valores de erro quadrático médio muito elevados na região definida por elevados valores tanto de x como de y. Este comportamento idêntico observado nos gráficos do viés e do erro quadrático médio está relacionado com o facto de o erro quadrático médio ser igual à soma da variância com o quadrado do viés. Uma vez que a variância apresenta uma variação muito idêntica em todos os cenários, havendo diferenças na variação do viés, é de esperar que haja uma relação idêntica observável entre a variação do viés e do erro quadrático médio.

Pode observar-se uma aproximação a zero por parte do erro quadrático médio de $\hat{F}_{12}(x,y)$ com a diminuição da percentagem de observações censuradas em cada um dos cenários de tamanho de amostra e de correlação. Registou-se também uma aproximação a zero por parte do erro quadrático médio com o aumento da dimensão da amostra, facto que sugere a verificação da

condição suficiente de consistência (1.3.9). Este facto pode ainda ser observado no gráfico da Figura A.7.

4.2.2 Estimador Kaplan-Meier pesado

Os gráficos da Figura B.1 e da Figura B.2 apresentam padrões de linhas de nível, bem como de intensidades de cor idênticos. Sugerindo que não existem diferenças significativas quanto ao viés das estimativas $\hat{F}_{12}(x,y)$ segundo o estimador KMW para cada um dos cenários de correlação simulados. Pelo que o que será escrito de seguida em relação ao viés de $\hat{F}_{12}(x,y)$ pode ser considerado independente do cenário de correlação. Assim para baixos valores dos quantis x e y o viés registado encontra-se muito próximo do zero. À medida que as variáveis x e y aumentam, o viés de $\hat{F}_{12}(x,y)$ diminui. Sendo esta variação linear quase perfeita. Pois a linhas de nível registadas nas caudas superiores são linhas rectas aparentemente perpendiculares à recta imaginária que une a origem (canto inferior esquerdo) com o ponto para o qual as variáveis x e y correspondem ao máximo valor representado no gráfico (canto superior direito). Portanto o estimador KMW em média subestima o verdadeiro valor $F_{12}(x,y)$.

Observa-se uma aproximação a zero por parte do viés quando há uma diminuição da percentagem de censura. Um aumento da dimensão da amostra é acompanhado de uma aproximação a zero por parte do viés. O que sugere a verificação da condição insuficiente de consistência (1.3.10).

O padrão de linhas de nível e intensidades de cor de desvio padrão também não é significativamente diferente entre ambos os cenários de correlação. Como se pode constatar pela observação dos gráficos da Figura B.3 e da Figura B.4. Até a amplitude das escalas de cor é igual entre os dois cenários de correlação. As linhas de nível de desvio padrão igual ou aproximado estão localizadas em regiões idênticas em ambos os gráficos. Em ambos os casos o desvio padrão aumenta com o aumento das variáveis x e y.

Verifica-se uma aproximação a zero por parte do desvio padrão com a diminuição da percentagem de censura. O desvio padrão também se aproxima de zero à medida que aumenta o tamanho da amostra, sugerindo a verificação da condição insuficiente de consistência (1.3.11).

Segundo os gráficos da Figura B.5 e da Figura B.6, não se observam diferenças significativas entre o erro quadrático médio de ambos os cenários de correlação. O que é de esperar, dado não

terem sido anteriormente observados padrões de variação muito distintos em relação ao viés e ao desvio padrão. Ambos os cenários de correlação evidenciam aumentos do erro quadrático médio com o aumento de ambas as variáveis x e y.

Tal como foi anteriormente observado em relação ao viés e ao desvio padrão, o erro quadrático médio das estimativas $\hat{F}_{12}(x,y)$ tende a aproximar-se de zero com a diminuição da percentagem de censura. Aumentos do tamanho de amostra são acompanhados de uma aproximação a zero por parte do erro quadrático médio, sugerindo a verificação da condição suficiente de consistência (1.3.9). O gráfico da Figura B.7 sugere a verificação da mesma condição de consistência.

4.2.3 Estimador Kaplan-Meier pesado pré-suavizado

O gráfico da Figura C.1 sugere que o estimador KMPW em média está a subestimar o verdadeiro valor de $F_{12}(x,y)$. Isto no cenário de correlação nula. Observam-se valores de viés próximos de zero para valores de x e y próximos da origem. Havendo uma tendência para o viés diminuir, afastando-se de zero à medida que os valores das variáveis x e y aumentam. No cenário de maior percentagem de censura e tamanho de amostra igual a 800, é evidente um aumento do viés seguido de um decréscimo

No cenário de correlação máxima o viés representado no gráfico da Figura C.2, permite verificar que o estimador KMPW também subestima em média o valor de $F_{12}(x,y)$. No entanto podem ser observadas algumas diferenças. Foram registados valores de viés próximos de zero nas caudas inferiores da função de distribuição Sendo que o viés no geral diminui à medida que aumentam os valores dos quantis x e y. A região de maiores valores de x e menores valores de y, próxima do canto inferior direito dos gráficos apresenta valores positivos de viés. Sendo observada uma tendência para o viés aumentar no sentido do canto inferior direito dos gráficos, que é mais evidente nos cenários de maior tamanho de amostra.

Em ambos os cenários de correlação observa-se uma tendência para o viés se aproximar de zero com a diminuição da percentagem de censura. No geral verifica-se uma aproximação a zero do viés à medida que aumenta o tamanho da amostra. Sendo a excepção o cenário de correlação máxima e maior percentagem de censura, onde na região inferior direita da função de distribuição se verifica um afastamento positivo de zero por parte do viés de $\hat{F}_{12}(x,y)$.

Não são evidentes diferenças significativas entre o gráfico da Figura C.3 e da Figura C.4. O que sugere que a correlação não tem influência no comportamento do desvio padrão. O desvio padrão tende a aumentar à medida que aumentam as variáveis x e y. Tende a diminuir com a diminuição da percentagem de censura. Tende a diminuir com o aumento do tamanho amostral, sugerindo a verificação da condição insuficiente de consistência (1.3.11).

O erro quadrático médio de $\hat{F}_{12}(x,y)$ é idêntico em relação ao dois cenários de correlação estudados, como sugere uma comparação do gráfico da Figura C.6 com o gráfico da Figura C.7. No cenário de correlação máxima a escala de intensidades de cor apresenta uma amplitude um pouco superior. Verifica-se um aumento do erro quadrático médio com o aumento das variáveis x e y. O erro quadrático médio aproxima-se de zero quando há uma diminuição da percentagem de censura. Havendo também uma aproximação a zero quando há um aumento da dimensão da amostra. O que sugere a validação da condição suficiente de consistência (1.3.9). A mesma tendência pode ser observada no gráfico da Figura C.7. Este comportamento sugere que a componente de variância contribui mais para o erro quadrático médio que a componente de viés. Uma vez que foi observado no cenário de correlação máxima e maior percentagem de censura, um aumento de viés com a dimensão da amostra na região inferior direita.

4.2.4 Estimador de Lin

Considerando os gráficos da Figura D.1 e da Figura D.2, pode ser constatado que o estimador de Lin em média sobrestima o verdadeiro valor de $F_{12}(x,y)$. A escala de intensidades de cor apresenta maior amplitude no cenário de correlação máxima. Sugerindo que o mesmo estimador sobrestima mais no cenário de correlação máxima. Verificam-se também linhas de nível de maior valor de viés no cenário de correlação máxima. Os valores de viés mais próximos de zero podem ser observados na região inferior esquerda definida por menores valores das variáveis x e y. Observa-se uma tendência para o viés aumentar à medida que aumentam os valores das variáveis x e y. No entanto no cenário de maior percentagem de censura o viés aumenta no sentido da região próxima do ponto $(2.5,2.5)$, diminuindo à medida que os quantis se afastam deste ponto.

Em ambos os cenários de correlação o viés aproxima-se de zero com a diminuição da percentagem de censura. No entanto quando o tamanho da amostra aumenta não é evidente uma aproximação a zero por parte do viés. Se esta aproximação existir então a taxa de variação do viés em relação ao tamanho da amostra deve ser muito baixa.

Os gráficos da Figura D.3 e Figura D.4 são muito idênticos em relação à disposição das linhas de nível, bem como em relação à disposição das intensidades de cor. A amplitude da escala de intensidades de cor é aproximadamente igual em ambos os gráficos. Estes factos sugerem que não existem diferenças significativas quanto ao comportamento do desvio padrão de $\hat{F}_{12}(x, y)$ em relação ao cenário de correlação. Observa-se uma tendência para o desvio padrão aumentar no sentido da região definida por baixos valores da variável x e valores muito altos da variável y. O desvio padrão tende a diminuir com a diminuição da percentagem de censura e com o aumento do tamanho da amostra. Sugerindo a validação da condição insuficiente de consistência (1.3.11).

Uma comparação entre os gráficos da Figura D.5 e da Figura D.6 permite constatar que o comportamento do erro quadrático médio não é muito diferente entre os dois cenários de correlação simulados. O padrão de variação do erro quadrático médio em relação aos quantis da função de distribuição é muito idêntico ao observado em relação ao desvio padrão. O que sugere que a componente de variância se reflecte mais no erro quadrático médio que a componente de viés. Como em todos os estimadores anteriormente analisados, verifica-se uma aproximação a zero do erro quadrático médio com a diminuição da percentagem de censura. A aproximação a zero do erro quadrático médio à medida que aumenta o tamanho da amostra sugere a validação da condição suficiente de consistência (1.3.9). A Figura D.7 apresenta uma representação gráfica alternativa que sugere a validação da mesma condição.

4.2.5 Eficiência relativa

De forma a facilitar a comparação dos quatro estimadores estudados relativamente ao erro quadrático médio; foram determinadas as eficiências relativas dos estimadores dois a dois para cada um dos cenários de correlação, censura e tamanho de amostra simulados. A partir destes resultados foram construídos os gráficos do Anexo E. As representações gráficas são mapas de intensidades de cor com linhas de nível relativas a valores iguais de eficiência relativa. De forma a facilitar a identificação do estimador mais eficiente, as regiões para as quais a eficiência relativa é superior à unidade estão pintadas com cores diferentes das cores utilizadas para as regiões para as quais a eficiência relativa é inferior à unidade. Assim as regiões para as quais a eficiência relativa é superior à unidade são pintadas com cores quentes em diferentes intensidades conforme o valor da eficiência relativa nessa região. Sendo as regiões para as quais a eficiência relativa é inferior à unidade pintadas com cores frias em diferentes intensidades conforme o respectivo valor da eficiência relativa. Sendo desta forma simplificadas as interpretações gráficas.

Segundo os gráficos da Figura E.7 e da Figura E.8 o estimador KMPW é mais eficiente que o estimador KMW em quase toda a região representada da função de distribuição bivariada. Sendo a excepção a região inferior direita no cenário de tamanho de amostra igual a 800, onde o estimador KMW é superior ao estimador KMPW. Este facto está relacionado com a anomalia observada em relação ao viés do estimador KMPW na mesma região. Reveja-se a subsecção 4.2.3. Apesar desta excepção, o estimador KMPW é superior em quase todo o plano representado. Nos restantes cenários de amostra o estimador KMPW revelou-se mais eficiente que o estimador KMW em todo o plano. O estimador KMPW torna-se mais eficiente que o estimador KMW à medida que aumentam os valores das variáveis x e y. Face a estes resultados, o estimador KMPW deve ser preferido em relação ao estimador KMW. Restando comparar os estimadores KMPW, CKM e LIN.

A observação dos gráficos da Figura E.11 e da Figura E.12 revela que a eficiência do estimador KMPW é superior à do estimador LIN na maior parte da região representada. No entanto para valores superiores da variável x, o estimador LIN revela-se mais eficiente que o estimador KMPW. Nos cenários de menor censura esta tendência não é bem evidente. Não se observam diferenças significativas entre os dois cenários de correlação simulados. Assim para valores inferiores da variável x, o estimador KMPW deve ser preferido em relação ao estimador LIN. Para valores superiores da variável x, o estimador LIN deve ser preferido em relação ao estimador KMPW.

A eficiência relativa do estimador KMPW em relação ao estimador CKM encontra-se representada nos gráficos da Figura E.3 e da Figura E.4. Em geral verifica-se que o estimador KMPW é mais eficiente para valores mais baixos da variável x. A partir de determinado valor da variável x o estimador CKM torna-se mais eficiente que o estimador KMPW.

O estimador LIN pode ser comparado com o estimador CKM em termos de eficiência relativa, recorrendo a uma análise dos gráficos da Figura E.5 e da Figura E.6. Os gráficos sugerem que o estimador CKM é mais eficiente para menores valores da variável x. Para valores superiores da variável x, o estimador LIN revela-se mais eficiente. Com excepção para os cenários de tamanho amostral maior. No entanto para tamanhos amostrais superiores todos os estimadores revelaram erros quadráticos médios muito baixos. Pelo que é preferível tomar uma decisão com base nos resultados observados em tamanhos amostrais inferiores.

Para finalizar, pode-se constatar que a eficiência relativa depende mais da variável x que da variável y. Uma vez que em quase todos os gráficos foram observados rectângulos coloridos verticais perpendiculares ao eixo das abcissas, onde a variável x se encontra representada. Apresentando cada um dos rectângulos, da esquerda ou da direita, cores quentes ou cores frias, que

indicam a qual dos sentidos de afastamento em relação à unidade a eficiência relativa se refere. Assim uma decisão simples baseada na eficiência relativa, seria preferir o estimador KMPW para baixos valores da variável x, para valores intermédios da mesma variável preferir o estimador CKM, e para valores superiores da mesma variável preferir o estimador LIN. Pelo que o estimador LIN deve ser preferido para estimar a função de distribuição marginal do segundo intervalo de tempo.

5 Conclusões

Neste trabalho foram apresentados quatro estimadores para a função de distribuição bivariada para tempos sequenciais na presença de censura pela direita. Todos os estimadores bivariados apresentados têm por base um estimador univariado conhecido por estimador Kaplan-Meier. De modo a proporcionar resultados numéricos e gráficos em relação aos quatro estimadores apresentados, foi desenvolvida uma extensão para o *software* estatístico R. Os quatro estimadores foram alvo de um estudo de simulação. O estudo de simulação detalhadamente descrito na secção 4 permitiu obter algumas conclusões relativamente a cada um dos estimadores.

Relativamente às propriedades desejáveis das funções de distribuição bivariadas, foi possível concluir que alguns dos estimadores não as garantem na totalidade. Sendo que a observação de algumas das propriedades pode depender da amostra. Assim o estimador Kaplan-Meier condicional não garante três propriedades entre as cinco possíveis. Os estimadores Kaplan-Meier pesado e Kaplan-Meier pesado pré-suavizado não garantem apenas uma das cinco propriedades. O estimador de Lin não garante quatro das cinco propriedades. Sendo que nenhum dos quatro estimadores estudados garante a propriedade $F(+\infty, +\infty) = 1$. Sendo desejável que os estimadores das funções de distribuição verifiquem as mesmas propriedades que as funções de distribuição teóricas. Saem a ganhar os estimadores Kaplan-Meier pesado e Kaplan-Meier pesado pré-suavizado dado que garantem o maior número de propriedades teóricas entre os quatro estimadores comparados. Foi ainda observado que os estimadores Kaplan-Meier pesado e Kaplan-Meier pesado pré-suavizado são idênticos e que o segundo apresenta menor variância relativamente ao primeiro.

Relativamente ao enviesamento foi possível constatar que uns estimadores em algumas situações, em média subestimam o verdadeiro valor da função de distribuição bivariada. Sendo que outros estimadores em média sobrestimam o verdadeiro valor da mesma função. O estimador Kaplan-Meier condicional em amostras de correlação nula, em média subestima o verdadeiro valor da função de distribuição bivariada. Em amostras cujos intervalos de tempo estão correlacionados, o estimador Kaplan-Meier condicional em média sobrestima o verdadeiro valor da função de distribuição bivariada. Em qualquer dos cenários de correlação e censura, o estimador Kaplan-Meier pesado e o estimador Kaplan-Meier pesado pré-suavizado em média subestimam o verdadeiro valor da função de distribuição bivariada. No entanto o estimador Kaplan-Meier pesado pré-suavizado manifesta uma tendência para sobrestimar a função de distribuição bivariada na cauda inferior direita. Por sua vez, o estimador de Lin em média sobrestima a função de distribuição bivariada, sobrestimando mais em amostras para as quais os intervalos de tempo estão correlacionados.

De estimador para estimador, o enviesamento apresenta diferentes comportamentos nas caudas da função de distribuição bivariada. Assim o estimador Kaplan-Meier condicional no cenário de independência manifesta um afastamento de zero por parte do viés no sentido da cauda superior direita. Enquanto no cenário de correlação máxima, o mesmo estimador registou um afastamento de zero em relação ao viés no sentido da cauda superior esquerda da função de distribuição bivariada. Os restantes três estimadores registaram afastamentos de zero por parte do viés no sentido da cauda superior direita da função de distribuição bivariada.

Em relação ao desvio padrão, foram registados padrões diferentes de variação nas caudas em relação ao estimador de Lin. O desvio padrão do estimador de Lin tende a aumentar no sentido da cauda superior esquerda da função de distribuição bivariada. Enquanto nos outros três estimadores o desvio padrão exibe uma tendência para aumentar no sentido da cauda superior direita da função de distribuição bivariada.

O estimador Kaplan-Meier condicional registou padrões de variação de erro quadrático médio distintos nos dois cenários de correlação simulados. No cenário de correlação nula o erro quadrático médio aumenta no sentido da cauda superior direita da função de distribuição bivariada; sendo que no cenário de correlação máxima aumenta no sentido da cauda superior esquerda. O erro quadrático médio do estimador de Lin também exibe uma tendência para aumentar no sentido da cauda superior esquerda da função de distribuição bivariada, neste caso em ambos os cenários de correlação. Os estimadores Kaplan-Meier pesado e Kaplan-Meier pesado pré-suavizado registaram aumentos do erro quadrático médio no sentido da cauda superior direita da função de distribuição bivariada.

Em todos os estimadores estudados foi observada uma aproximação a zero do enviesamento, do desvio padrão e do erro quadrático médio quando houve uma diminuição da percentagem de censura. Os quatro estimadores estudados manifestaram uma aproximação a zero do desvio padrão e do erro quadrático médio acompanhada de um aumento da dimensão da amostra. No entanto tal aproximação com o aumento do tamanho da amostra nem sempre foi evidente em relação ao enviesamento. Tendo sido evidente apenas em relação ao estimador Kaplan-Meier condicional em amostras cujos intervalos de tempo são independentes. É importante referir que a aproximação a zero por parte do erro quadrático médio à medida que aumenta o tamanho da amostra apenas sugere a validação da condição suficiente consistência. Não permitindo concluir que um determinado estimador é consistente. Pois nenhum método gráfico permite tirar essa conclusão.

A comparação dos quatro estimadores entre si em termos da eficiência relativa permitiu inicialmente seleccionar o estimador Kaplan-Meier pesado pré-suavizado em detrimento do estimador Kaplan-Meier pesado. Em seguida a comparação do estimador Kaplan-Meier pesado pré-suavizado, do estimador Kaplan-Meier condicional, e do estimador de Lin, levou à conclusão que a eficiência relativa entre estes três estimadores dois a dois depende da variável x, que se refere ao primeiro intervalo de tempo. Assim foi possível concluir que para valores mais baixos da variável x, o estimador Kaplan-Meier pesado pré-suavizado é o mais eficiente entre os três estimadores em comparação. Para valores intermédios da variável x o estimador Kaplan-Meier condicional é o mais eficiente. Para valores superiores da variável x destaca-se o estimador de Lin como o mais eficiente. Pelo que o estimador de Lin se revela mais eficiente na estimativa da função de distribuição marginal do segundo intervalo de tempo.

Considerando as garantias em relação às propriedades da função de distribuição bivariada e a eficiência relativa em simultâneo, pode-se concluir que o estimador mais desejável é o estimador Kaplan-Meier pesado pré-suavizado.

Esta página foi intencionalmente deixada em branco.

Anexo A Estimador Kaplan-Meier condicional

Tabela A.1: Enviesamento de $\widehat{F}_{12}(x,y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita.

		$\alpha = 0$				$\alpha = 1$			
	Y	0.2231	0.5108	0.9163	1.6094	0.2231	0.5108	0.9163	1.6094
	x								
$C \sim U[0,4]$	0.2231	0.11877	-0.184271	0.039722	-0.780763	0.46965	0.9363	0.77021	-1.110113
	0.5108	0.31372	-0.034404	0.530963	-0.018024	0.96987	1.7148	1.60922	-0.071925
	0.9163	0.43170	0.058303	0.647391	0.111059	2.60263	4.5782	4.64682	2.389977
	1.6094	-0.28917	-0.593437	-0.333192	-0.856723	7.80302	12.4596	13.18540	8.629928
	$n = 50$								
	0.2231	-0.20034	-0.34649	-0.44602	-0.237546	0.19258	0.21964	0.087955	-0.37785
	0.5108	-0.14947	-0.52160	-0.22029	0.036627	0.81711	1.34557	1.540752	1.16243
	0.9163	-0.24991	-0.47242	0.23824	0.493575	2.71279	4.17077	4.323562	3.28648
	1.6094	-0.44859	-0.74597	0.22609	0.406282	7.75228	12.23943	12.646230	9.23350
	$n = 100$								
	0.2231	-0.145385	-0.072459	0.17995	0.24604	0.19258	0.21964	0.087955	-0.37785
	0.5108	-0.200762	-0.053334	0.29844	0.38039	0.81711	1.34557	1.540752	1.16243
	0.9163	-0.050097	0.254858	0.42100	0.20061	2.71279	4.17077	4.323562	3.28648
	1.6094	-0.087924	0.176399	0.42224	0.20893	7.75228	12.23943	12.646230	9.23350
	$n = 200$								
	0.2231	0.049456	0.17216	0.195110	0.177617	-0.21304	-0.32287	-0.26334	-0.34367
	0.5108	0.225746	0.12289	0.161403	0.104623	0.37486	0.64145	0.64396	0.43669
	0.9163	0.202001	0.16135	-0.009111	0.087799	2.28782	3.55204	3.72709	2.86834
	1.6094	0.224313	0.27370	-0.016976	0.096507	7.62425	11.96464	12.57394	9.45116
	$n = 400$								
$C \sim U[0,3]$	0.2231	0.07515	-0.40174	-0.43970	-2.23666	0.01205	-0.15326	-0.4843	-3.2603
	0.5108	0.07376	-0.01063	0.33126	-0.57780	1.29921	1.89040	2.4391	-0.2193
	0.9163	-0.08100	0.20984	0.62913	-0.16169	4.33889	6.08457	6.5349	3.1156
	1.6094	-0.22594	-0.02901	0.27949	-0.63010	11.8171	17.77896	19.6212	13.9123
	$n = 50$								
	0.2231	0.16008	0.106228	0.13962	-0.27226	0.15255	-0.29525	-0.014864	-0.98326
	0.5108	0.10719	0.052657	0.29725	0.21420	0.94779	1.21937	1.600822	1.18659
	0.9163	0.46418	0.275293	0.23151	0.24430	3.47179	5.23757	6.344859	4.67998
	1.6094	0.49309	0.440719	0.30528	0.20176	11.01872	17.08305	19.675171	15.79618
	$n = 100$								
	0.2231	-0.156217	-0.148194	-0.14399	-0.14236	-0.04686	0.15116	0.32616	0.046488
	0.5108	-0.027277	-0.197454	0.13580	0.23433	0.69588	1.35945	1.75663	1.544016
	0.9163	0.059506	0.001014	0.32464	0.74276	3.01595	5.39362	6.24912	4.891408
	1.6094	0.036410	-0.192298	0.23253	0.63987	10.51300	17.48421	19.72548	15.957647
	$n = 200$								
	0.2231	0.19012	0.087995	0.30747	0.24871	0.01117	0.29969	0.1098	0.05347
	0.5108	0.10178	0.075911	0.20832	0.20355	0.94721	1.38828	1.4918	1.19265
	0.9163	0.15710	-0.054610	0.18084	0.44802	3.57191	5.38416	5.8443	4.65529
	1.6094	0.21125	-0.120136	0.28377	0.58377	11.21911	17.58371	19.2605	15.62439
	$n = 400$								

Tabela A.2: Desvio padrão de $\widehat{F}_{12}(x,y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita.

		$\alpha = 0$				$\alpha = 1$			
	Y	0.2231	0.5108	0.9163	1.6094	0.2231	0.5108	0.9163	1.6094
	x								
$C \sim U[0,4]$	0.2231	28.454	39.915	48.573	55.460	36.438	48.298	54.841	58.120
	0.5108	40.126	55.153	65.790	73.013	48.714	62.757	70.446	73.475
	0.9163	49.630	66.845	76.548	83.287	56.609	71.750	78.751	80.547
	1.6094	57.872	75.749	84.276	88.019	62.587	78.784	84.336	83.819
	$n = 50$								
	0.2231	20.075	27.775	33.879	38.830	25.504	33.355	37.968	40.215
	0.5108	28.823	39.179	46.311	51.548	33.933	43.781	48.988	51.546
	0.9163	34.993	47.127	53.976	57.690	39.312	49.542	54.553	56.258
	1.6094	40.797	53.582	59.438	60.512	43.486	53.927	58.090	58.137
	$n = 100$								
	0.2231	14.190	19.840	24.343	27.520	25.504	33.355	37.968	40.215
	0.5108	19.935	27.511	32.811	36.235	33.933	43.781	48.988	51.546
	0.9163	24.709	33.303	38.518	40.964	39.312	49.542	54.553	56.258
	1.6094	28.667	37.933	42.002	42.613	43.486	53.927	58.090	58.137
	$n = 200$								
	0.2231	10.084	14.128	17.069	19.488	12.707	16.785	19.062	20.241
	0.5108	14.223	19.458	23.040	25.626	17.022	22.033	24.514	25.882
	0.9163	17.699	23.605	27.009	28.775	19.758	25.279	27.472	28.540
	1.6094	20.466	26.672	29.578	29.859	21.808	27.542	29.545	29.443
	$n = 400$								
$C \sim U[0,3]$	0.2231	28.883	40.724	50.046	58.000	36.388	48.269	55.161	58.336
	0.5108	40.792	57.009	67.903	76.684	48.629	64.064	72.666	76.635
	0.9163	50.618	68.457	80.363	88.358	57.179	73.655	82.710	87.185
	1.6094	60.247	79.647	91.224	98.564	64.570	81.803	90.834	96.739
	$n = 50$								
	0.2231	20.209	28.332	34.836	40.136	25.694	33.648	38.286	41.080
	0.5108	28.380	39.217	47.543	54.070	34.577	44.335	49.336	53.642
	0.9163	35.435	47.165	56.108	62.364	40.488	50.958	56.675	60.856
	1.6094	41.722	54.715	63.239	69.238	45.976	57.181	62.650	66.877
	$n = 100$								
	0.2231	14.512	20.215	24.599	28.221	18.306	24.047	27.199	28.850
	0.5108	20.451	27.821	33.389	37.736	24.282	31.802	35.609	37.392
	0.9163	25.146	34.160	39.770	43.574	28.404	36.613	40.348	42.448
	1.6094	30.027	40.027	45.250	48.720	31.911	40.572	44.438	46.748
	$n = 200$								
	0.2231	10.156	14.189	17.239	19.703	12.753	16.819	19.039	20.436
	0.5108	14.432	19.952	23.613	26.527	16.860	22.150	24.803	26.463
	0.9163	17.807	24.241	28.121	31.013	19.925	25.703	28.350	30.096
	1.6094	21.045	28.133	31.704	34.167	22.564	28.731	31.226	32.909
	$n = 400$								

Tabela A.3: Erro quadrático médio de $\widehat{F}_{12}(x,y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita.

		$\alpha = 0$				$\alpha = 1$			
	Y	0.2231	0.5108	0.9163	1.6094	0.2231	0.5108	0.9163	1.6094
	x								
$C \sim U[0,4]$	0.2231	0.80956	1.5931	2.3591	3.0761	1.3278	2.3334	3.0078	3.3788
	0.5108	1.61000	3.0416	4.3281	5.3304	2.3738	3.9410	4.9647	5.3980
	0.9163	2.46303	4.4678	5.8594	6.9361	3.2110	5.1685	6.2226	6.4929
	1.6094	3.34893	5.7377	7.1018	7.7473	3.9776	6.3616	7.2858	7.0995
	$n = 50$								
	0.2231	0.40299	0.77152	1.1479	1.5077	0.65041	1.1125	1.4414	1.6172
	0.5108	0.83070	1.53511	2.1446	2.6569	1.15201	1.9184	2.4019	2.6581
	0.9163	1.22444	2.22100	2.9131	3.3281	1.55266	2.4716	2.9944	3.1754
	1.6094	1.66444	2.87126	3.5325	3.6615	1.95093	3.0576	3.5341	3.4649
	$n = 100$								
	0.2231	0.20136	0.39357	0.59254	0.75734	0.65041	1.1125	1.4414	1.6172
	0.5108	0.39740	0.75680	1.07656	1.31297	1.15201	1.9184	2.4019	2.6581
	0.9163	0.61047	1.10907	1.48364	1.67796	1.55266	2.4716	2.9944	3.1754
	1.6094	0.82172	1.43879	1.76420	1.81570	1.95093	3.0576	3.5341	3.4649
	$n = 200$								
	0.2231	0.10169	0.19961	0.29138	0.37976	0.16150	0.28182	0.36341	0.40976
	0.5108	0.20232	0.37859	0.53083	0.65661	0.28986	0.48580	0.60129	0.66999
	0.9163	0.31325	0.55718	0.72941	0.82790	0.39559	0.65157	0.76853	0.82267
	1.6094	0.41887	0.71139	0.87477	0.89150	0.53367	0.90165	1.03092	0.95612
	$n = 400$								
$C \sim U[0,3]$	0.2231	0.83415	1.6585	2.5045	3.3686	1.3239	2.3297	3.0426	3.4134
	0.5108	1.66386	3.2498	4.6105	5.8801	2.3662	4.1074	5.2858	5.8723
	0.9163	2.56189	4.6860	6.4579	7.8064	3.2879	5.4615	6.8829	7.6101
	1.6094	3.62941	6.3430	8.3210	9.7144	4.3085	7.0071	8.6350	9.5511
	$n = 50$								
	0.2231	0.40841	0.80262	1.2135	1.6108	0.66012	1.1321	1.4657	1.6883
	0.5108	0.80536	1.53784	2.2602	2.9234	1.19635	1.9669	2.4364	2.8786
	0.9163	1.25570	2.22435	3.1478	3.8889	1.65114	2.6239	3.2519	3.7250
	1.6094	1.74083	2.99358	3.9989	4.7935	2.23500	3.5611	4.3117	4.7216
	$n = 100$								
	0.2231	0.21060	0.40864	0.6051	0.79636	0.33507	0.57823	0.73984	0.83225
	0.5108	0.41821	0.77397	1.1147	1.42390	0.59003	1.01309	1.27094	1.40040
	0.9163	0.63227	1.16681	1.5816	1.89909	0.81582	1.36947	1.66687	1.82554
	1.6094	0.90154	1.60208	2.0474	2.37382	1.12875	1.95159	2.36365	2.43983
	$n = 200$								
	0.2231	0.10317	0.20130	0.29726	0.38823	0.16262	0.28295	0.36245	0.41759
	0.5108	0.20828	0.39804	0.55756	0.70367	0.28513	0.49251	0.61735	0.70162
	0.9163	0.31708	0.58757	0.79073	0.96191	0.40974	0.68956	0.83780	0.92733
	1.6094	0.44291	0.79140	1.00513	1.16759	0.63493	1.13457	1.34592	1.32701
	$n = 400$								

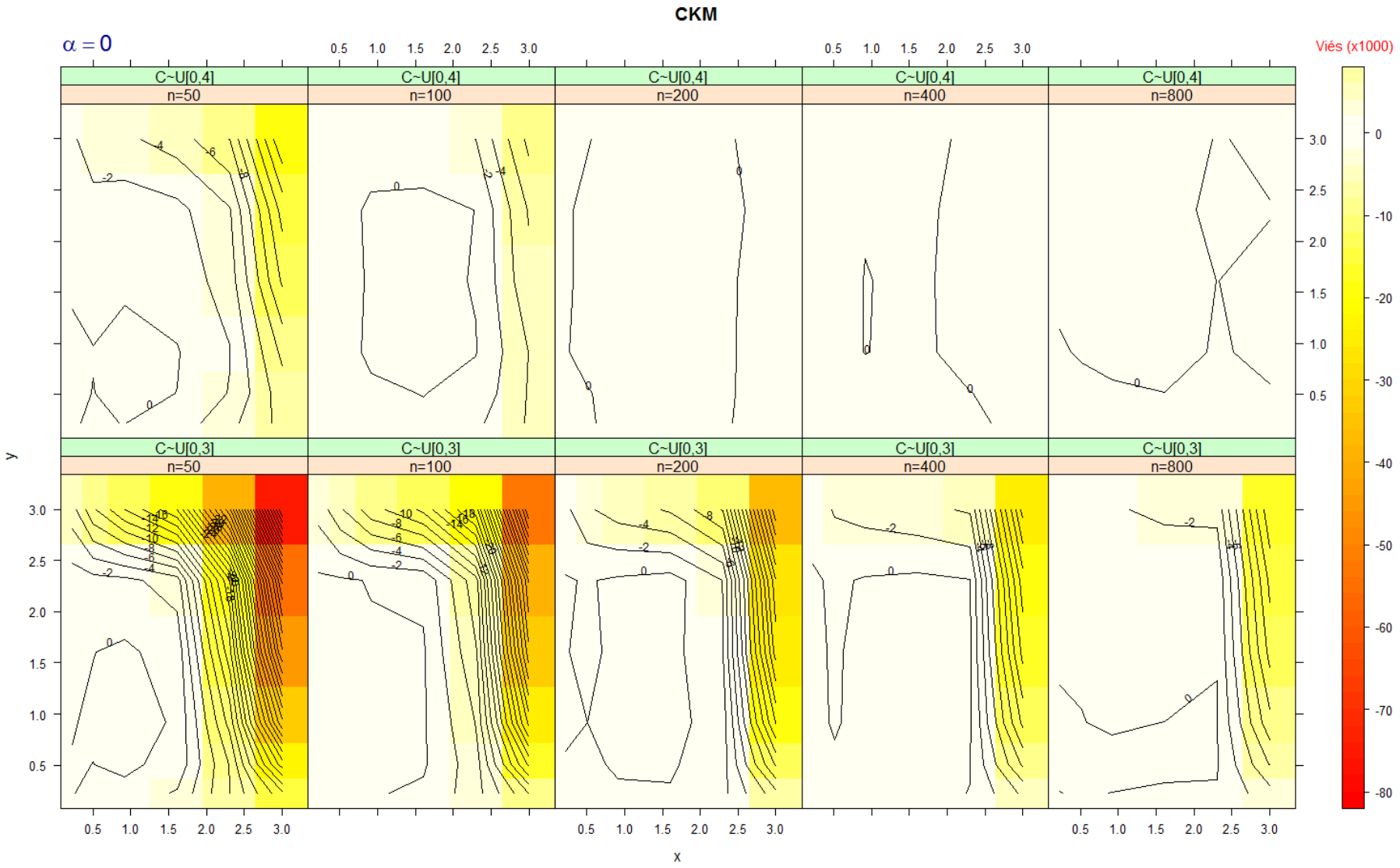


Figura A.1: Enviesamento de $\widehat{F}_{12}(x, y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 0$).

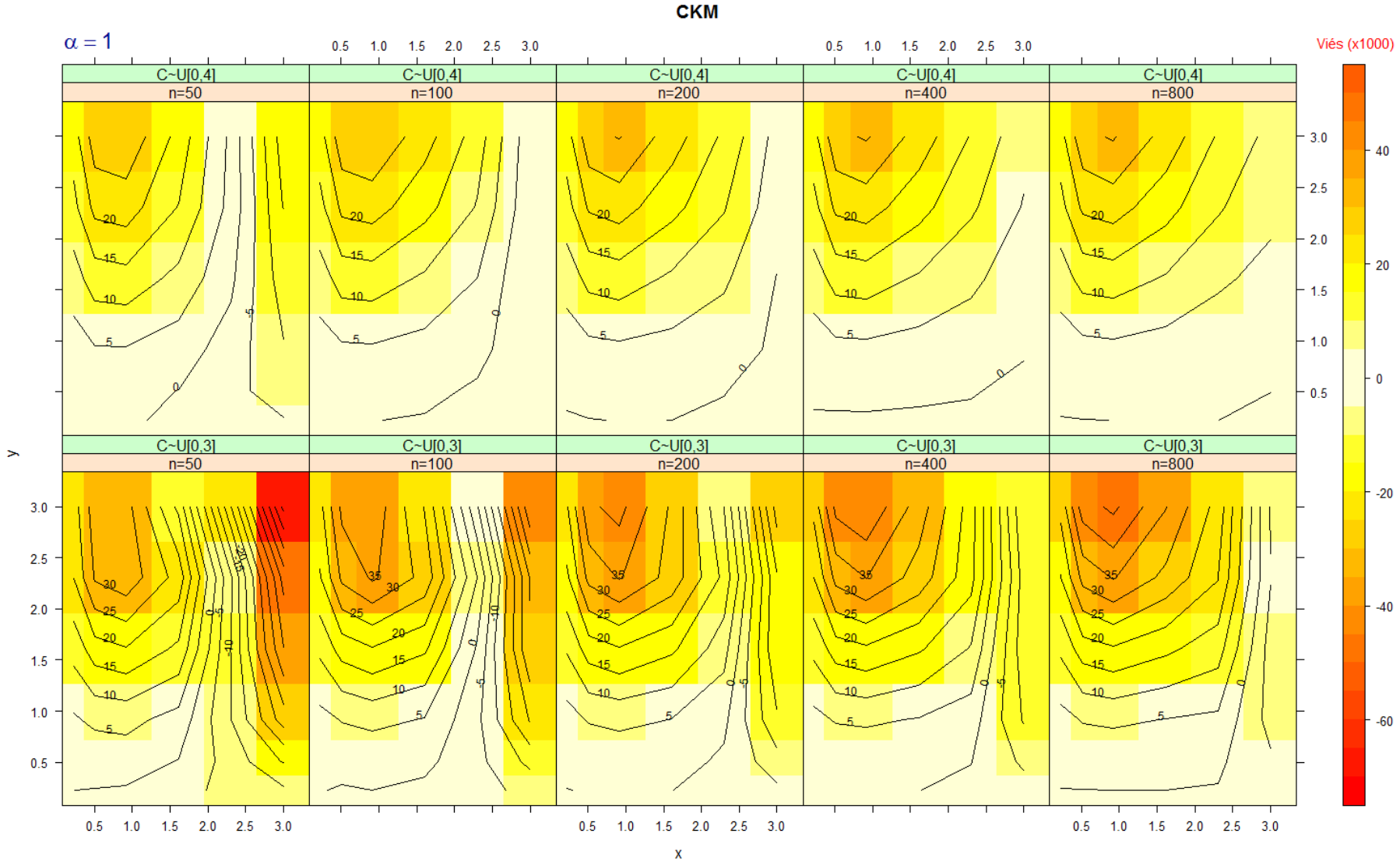


Figura A.2: Enviesamento de $\widehat{F}_{12}(x, y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 1$).

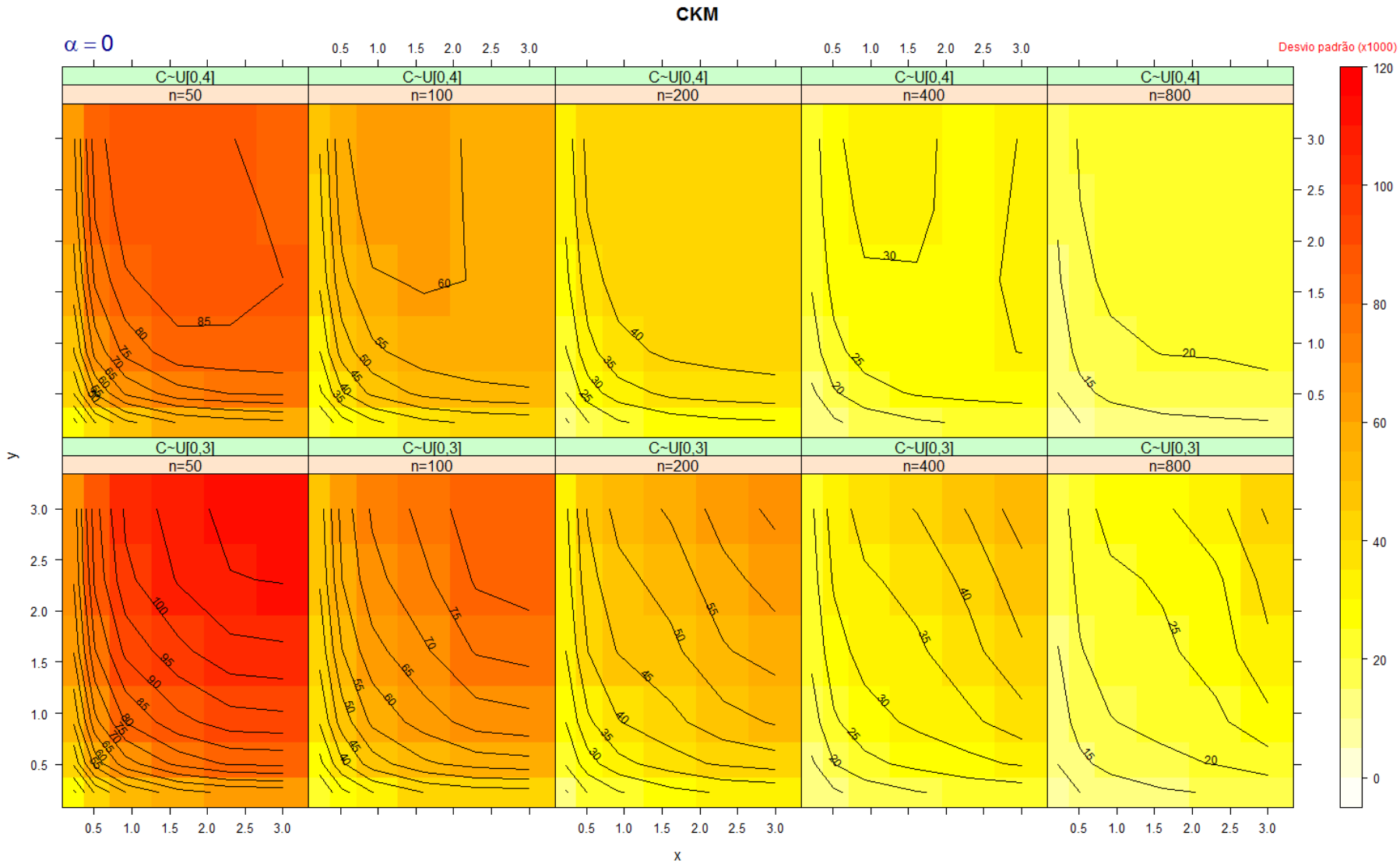


Figura A.3: Desvio padrão de $\widehat{F}_{12}(x, y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 0$).

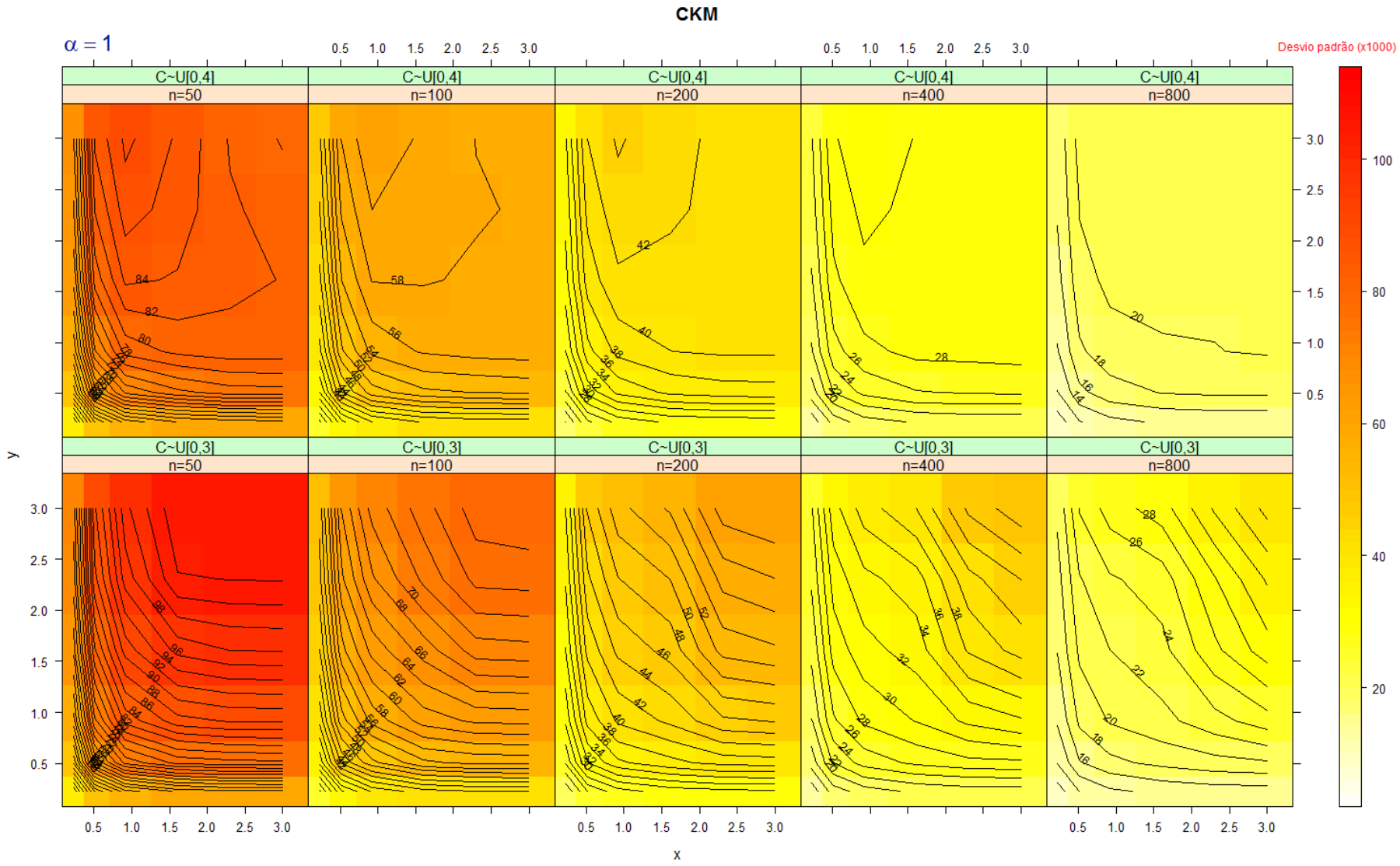


Figura A.4: Desvio padrão de $\widehat{F}_{12}(x,y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 1$).

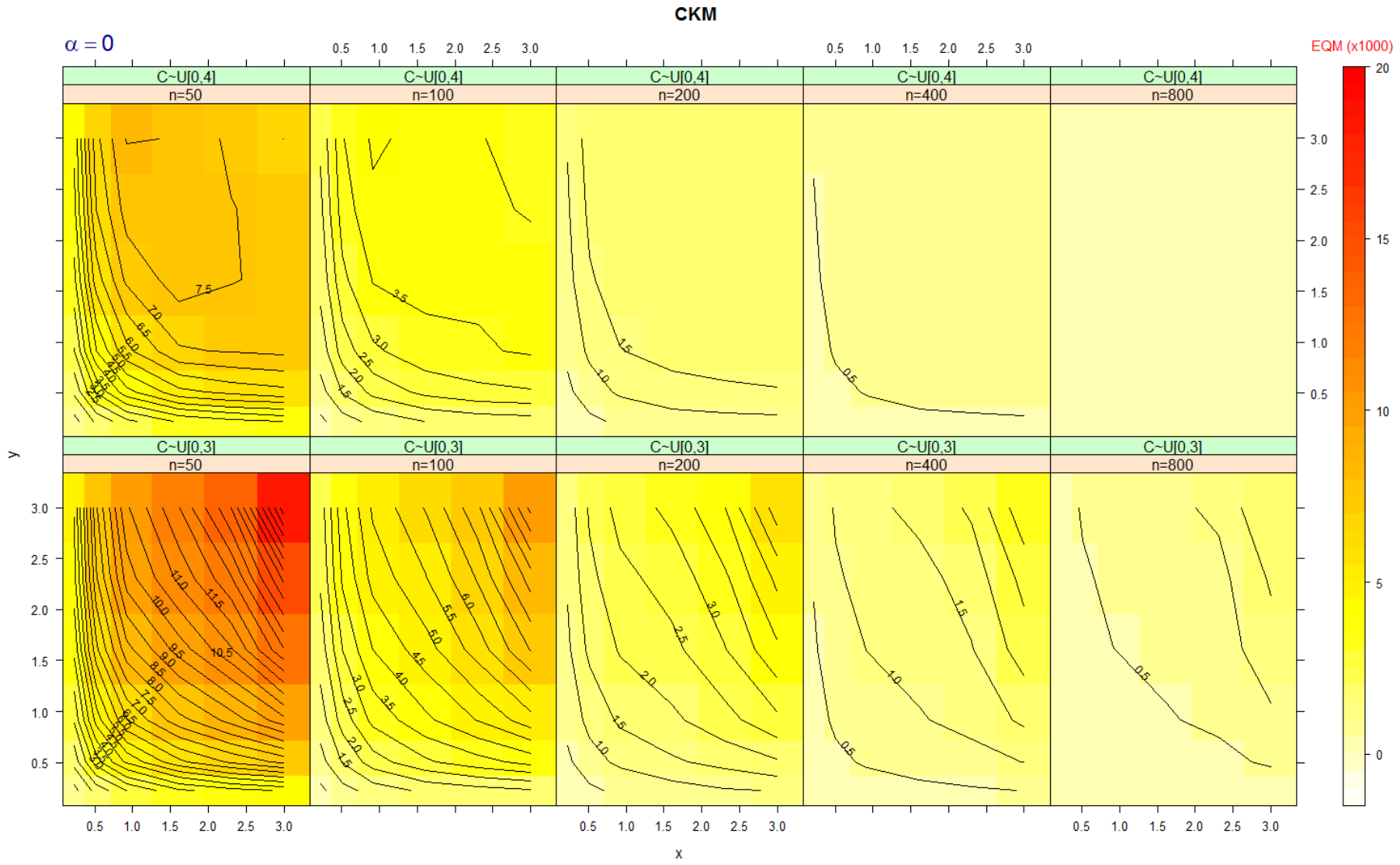


Figura A.5: Erro quadrático médio de $\widehat{F}_{12}(x, y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 0$).

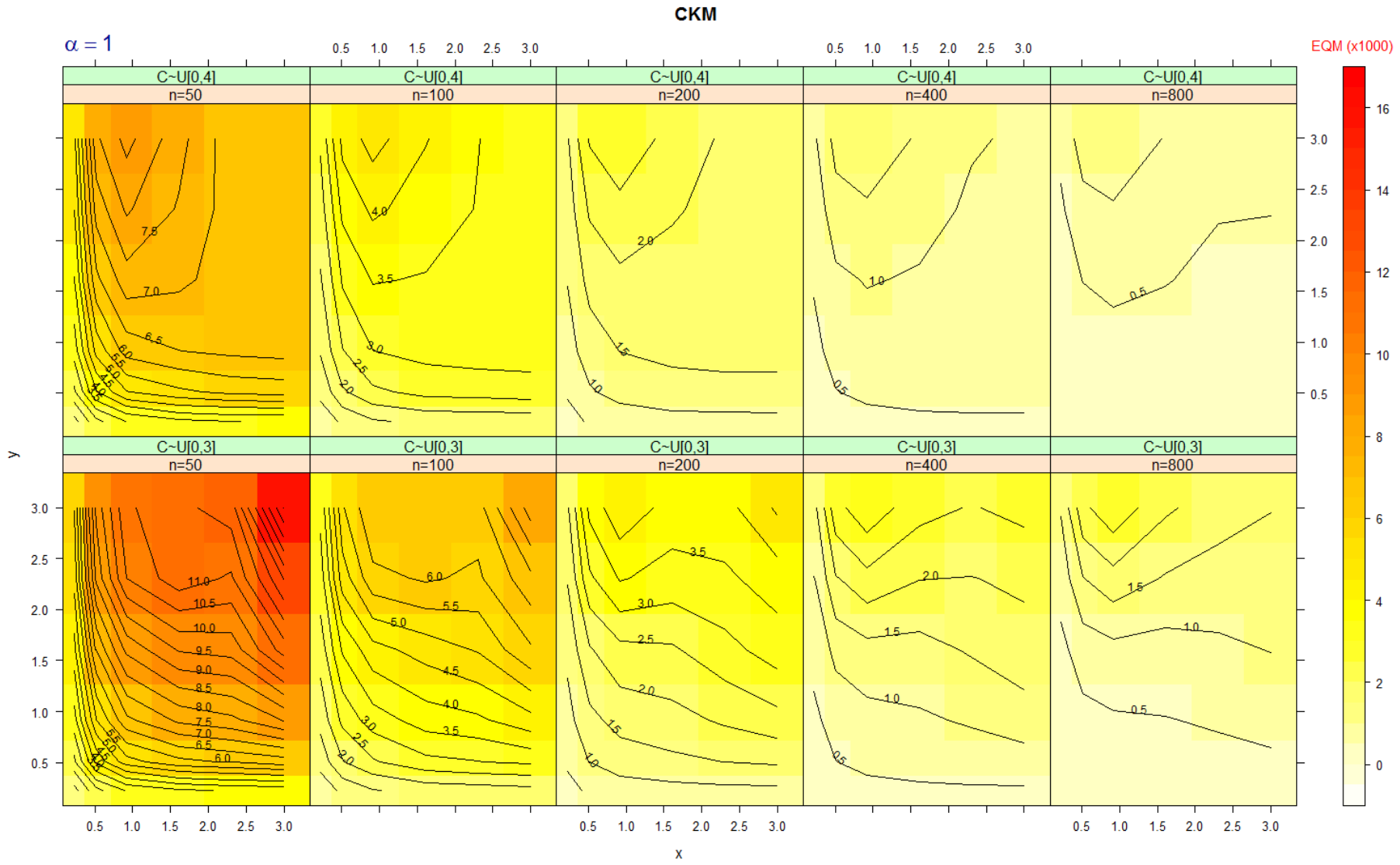


Figura A.6: Erro quadrático médio de $\widehat{F}_{12}(x, y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 1$).

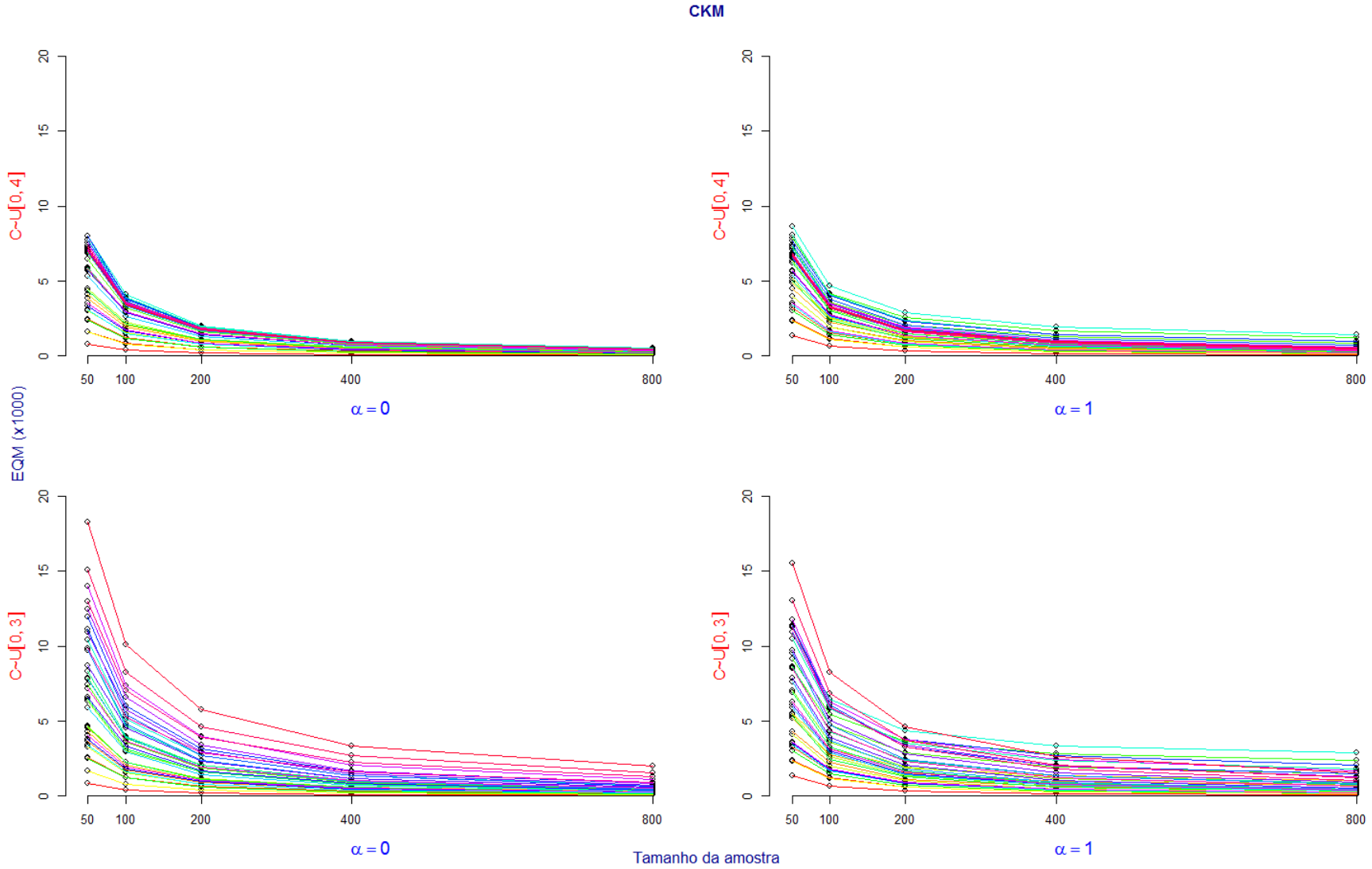


Figura A.7: Erro quadrático médio de $\widehat{F}_{12}(x, y)$ (x1000) versus o tamanho da amostra.

Anexo B Estimador Kaplan-Meier pesado

Tabela B.1: Enviesamento de $\widehat{F}_{12}(x,y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita.

		$\alpha = 0$				$\alpha = 1$			
	Y	0.2231	0.5108	0.9163	1.6094	0.2231	0.5108	0.9163	1.6094
	x								
$C \sim U[0,4]$	0.2231	0.11879	-0.12911	0.047105	0.133792	0.40831	0.79501	0.76289	0.56226
	0.5108	0.31464	-0.03643	0.512718	0.106239	0.34046	0.78642	0.57240	0.19521
	0.9163	0.45079	0.04222	0.508607	-0.063769	0.13996	0.73483	0.57211	-0.27199
	1.6094	-0.47769	-0.64835	-0.378363	-1.188151	0.16954	0.48202	0.59270	-0.80285
	$n = 50$								
	0.2231	-0.20675	-0.35857	-0.43292	-0.209700	0.164215	0.14422	0.062756	0.080874
	0.5108	-0.16784	-0.56580	-0.17245	0.048636	0.272669	0.50841	0.719859	0.608278
	0.9163	-0.28763	-0.51386	0.32871	0.603481	0.300876	0.39197	0.321761	0.116264
	1.6094	-0.54560	-0.91752	0.25394	0.483735	0.038208	0.17506	-0.261971	-0.527574
	$n = 100$								
	0.2231	-0.14897	-0.05638	0.209886	0.308720	0.164215	0.14422	0.062756	0.080874
	0.5108	-0.20385	-0.04469	0.275107	0.382383	0.272669	0.50841	0.719859	0.608278
	0.9163	-0.03291	0.30022	0.386011	0.142617	0.300876	0.39197	0.321761	0.116264
	1.6094	-0.03293	0.21757	0.365895	0.099851	0.038208	0.17506	-0.261971	-0.527574
	$n = 200$								
	0.2231	0.050086	0.18605	0.232181	0.230705	-0.286584	-0.467288	-0.44389	-0.53138
	0.5108	0.228673	0.10750	0.171914	0.140395	-0.211587	-0.299196	-0.38597	-0.33658
	0.9163	0.191396	0.14193	-0.012877	0.130495	-0.126386	-0.293699	-0.38226	-0.24395
	1.6094	0.192197	0.23068	-0.140637	-0.076618	0.021076	-0.009593	-0.20049	-0.10385
	$n = 400$								
$C \sim U[0,3]$	0.2231	0.09448	-0.38515	-0.22719	-0.39101	-0.057731	-0.31703	-0.18472	-0.095824
	0.5108	0.07128	0.05563	0.38044	0.33862	0.444117	0.50332	0.97630	0.587722
	0.9163	-0.13907	0.36676	0.54033	0.20269	1.129773	0.86679	0.79382	-0.059790
	1.6094	-0.17264	0.12763	0.27987	-5.12479	0.977814	0.38651	1.11308	-3.948700
	$n = 50$								
	0.2231	0.18916	0.159877	0.21300	0.044538	0.095558	-0.46100	-0.209555	0.071969
	0.5108	0.10309	0.018903	0.23388	-0.053396	0.138078	-0.14493	0.092580	0.403299
	0.9163	0.50151	0.332415	0.26538	0.207595	0.054108	-0.29449	0.038903	-0.573513
	1.6094	0.53315	0.589154	0.64914	-2.129035	-0.016546	-0.76091	-0.460355	-4.870312
	$n = 100$								
	0.2231	-0.16430	-0.18490	-0.18507	-0.14821	-0.12158	0.089890	0.33497	0.441099
	0.5108	-0.01831	-0.21233	0.14766	0.23085	-0.09906	0.120702	0.38995	0.569885
	0.9163	0.07142	-0.04711	0.23346	0.58653	-0.41334	-0.000713	0.35240	0.047017
	1.6094	0.05391	-0.24840	0.19910	-2.04827	-0.50349	-0.094441	0.44468	-2.108356
	$n = 200$								
	0.2231	0.192080	0.07594	0.271307	0.28217	-0.087639	0.130058	-0.114597	-0.202368
	0.5108	0.092501	0.07665	0.129097	0.14254	0.169343	0.105417	0.078171	0.044104
	0.9163	0.136525	-0.08794	0.079346	0.39678	0.170709	-0.065942	-0.116082	-0.100935
	1.6094	0.211952	-0.13388	0.304163	-0.66956	0.219970	-0.011106	-0.149406	-2.391015
	$n = 400$								

Tabela B.2: Desvio padrão de $\widehat{F}_{12}(x,y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita.

		$\alpha = 0$				$\alpha = 1$			
	Y	0.2231	0.5108	0.9163	1.6094	0.2231	0.5108	0.9163	1.6094
	x								
$C \sim U[0,4]$	0.2231	28.472	40.195	49.496	58.156	36.420	48.538	55.790	60.869
	0.5108	40.154	55.396	66.696	76.185	48.644	63.191	71.915	78.088
	0.9163	49.891	67.482	77.792	87.256	56.194	71.907	80.047	85.868
	1.6094	58.570	77.051	86.516	93.439	61.284	78.852	86.661	91.669
	$n = 50$								
	0.2231	20.098	27.963	34.608	41.083	25.550	33.685	39.073	42.656
	0.5108	28.850	39.387	47.153	54.396	33.901	44.225	50.349	54.884
	0.9163	35.064	47.403	54.910	60.765	39.044	49.768	55.879	60.238
	1.6094	41.219	54.328	61.095	64.238	42.498	53.843	59.774	63.598
	$n = 100$								
	0.2231	14.204	19.973	24.811	29.109	25.550	33.685	39.073	42.656
	0.5108	19.937	27.649	33.462	38.232	33.901	44.225	50.349	54.884
	0.9163	24.791	33.498	39.134	43.154	39.044	49.768	55.879	60.238
	1.6094	29.041	38.559	43.010	45.246	42.498	53.843	59.774	63.598
	$n = 200$								
	0.2231	10.099	14.207	17.414	20.520	12.717	16.919	19.536	21.358
	0.5108	14.245	19.608	23.598	27.180	16.997	22.165	25.075	27.403
	0.9163	17.756	23.762	27.498	30.322	19.623	25.348	28.107	30.347
	1.6094	20.660	27.064	30.378	31.870	21.349	27.486	30.387	31.974
	$n = 400$								
$C \sim U[0,3]$	0.2231	28.919	40.998	51.281	61.803	36.455	48.841	56.961	63.135
	0.5108	40.927	57.635	69.687	82.256	48.562	64.742	75.070	83.693
	0.9163	50.991	69.369	82.719	95.907	56.778	73.949	85.054	96.306
	1.6094	62.108	82.865	97.692	112.689	63.067	82.272	96.647	115.439
	$n = 50$								
	0.2231	20.241	28.545	35.726	42.963	25.808	34.141	39.709	44.563
	0.5108	28.463	39.552	48.958	58.044	34.579	44.862	51.210	58.406
	0.9163	35.774	47.746	57.786	67.706	40.190	51.429	58.794	67.262
	1.6094	42.998	56.808	67.517	82.659	44.781	57.702	66.481	82.098
	$n = 100$								
	0.2231	14.549	20.352	25.195	30.305	18.365	24.419	28.299	31.254
	0.5108	20.526	28.009	34.178	40.444	24.261	32.170	37.029	41.103
	0.9163	25.325	34.573	40.947	47.268	28.191	36.857	41.795	47.284
	1.6094	30.830	41.535	48.266	58.799	31.103	41.044	47.610	59.673
	$n = 200$								
	0.2231	10.186	14.345	17.804	21.404	12.770	17.054	19.779	22.037
	0.5108	14.470	20.146	24.336	28.777	16.849	22.394	25.647	28.751
	0.9163	17.919	24.544	28.906	33.792	19.801	25.951	29.364	33.467
	1.6094	21.563	29.130	33.596	42.386	22.090	29.164	33.323	42.781
	$n = 400$								

Tabela B.3: Erro quadrático médio de $\widehat{F}_{12}(x,y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita.

		$\alpha = 0$				$\alpha = 1$			
	Y	0.2231	0.5108	0.9163	1.6094	0.2231	0.5108	0.9163	1.6094
	x								
$C \sim U[0,4]$	0.2231	0.81057	1.6155	2.4496	3.3819	1.3264	2.3564	3.1128	3.7050
	0.5108	1.61231	3.0684	4.4482	5.8037	2.3661	3.9933	5.1715	6.0972
	0.9163	2.48905	4.5534	6.0513	7.6128	3.1575	5.1706	6.4071	7.3726
	1.6094	3.43038	5.9367	7.4845	8.7313	3.7554	6.2172	7.5097	8.4031
	$n = 50$								
	0.2231	0.40393	0.7820	1.1978	1.6877	0.65279	1.1346	1.5266	1.8194
	0.5108	0.83226	1.5515	2.2233	2.9586	1.14924	1.9559	2.5352	3.0124
	0.9163	1.22944	2.2471	3.0149	3.6924	1.52440	2.4767	3.1222	3.6283
	1.6094	1.69912	2.9521	3.7323	4.1264	1.80588	2.8988	3.5727	4.0446
	$n = 100$								
	0.2231	0.20175	0.39888	0.61557	0.84736	0.65279	1.1346	1.5266	1.8194
	0.5108	0.39748	0.76439	1.11965	1.46170	1.14924	1.9559	2.5352	3.0124
	0.9163	0.61451	1.12211	1.53144	1.86214	1.52440	2.4767	3.1222	3.6283
	1.6094	0.84327	1.48669	1.84978	2.04699	1.80588	2.8988	3.5727	4.0446
	$n = 200$								
	0.2231	0.10199	0.20185	0.30328	0.42109	0.16179	0.28645	0.38180	0.45641
	0.5108	0.20294	0.38445	0.55682	0.73872	0.28892	0.49133	0.62886	0.75097
	0.9163	0.31529	0.56461	0.75605	0.91934	0.38505	0.64257	0.79009	0.92091
	1.6094	0.42684	0.73245	0.92274	1.01558	0.45573	0.75541	0.92333	1.02225
	$n = 400$								
$C \sim U[0,3]$	0.2231	0.83623	1.6808	2.6296	3.8193	1.3288	2.3853	3.2442	3.9856
	0.5108	1.67487	3.3214	4.8559	6.7655	2.3582	4.1913	5.6359	7.0042
	0.9163	2.59989	4.8118	6.8421	9.1973	3.2247	5.4687	7.2342	9.2739
	1.6094	3.85699	6.8660	9.5429	12.7238	3.9780	6.7681	9.3410	13.3403
	$n = 50$								
	0.2231	0.40968	0.81476	1.2763	1.8456	0.66598	1.1657	1.5767	1.9857
	0.5108	0.81010	1.56420	2.3967	3.3688	1.19559	2.0124	2.6222	3.4111
	0.9163	1.27993	2.27956	3.3390	4.5836	1.61505	2.6448	3.4564	4.5240
	1.6094	1.84893	3.22720	4.5585	6.8364	2.00512	3.3298	4.4196	6.7631
	$n = 100$								
	0.2231	0.21167	0.41418	0.63477	0.9183	0.33724	0.59622	0.80089	0.97691
	0.5108	0.42129	0.78450	1.16807	1.6356	0.58852	1.03483	1.37119	1.68965
	0.9163	0.64130	1.19517	1.67651	2.2344	0.79482	1.35827	1.74675	2.23551
	1.6094	0.95039	1.72508	2.32944	3.4611	0.96753	1.68446	2.26671	3.56496
	$n = 200$								
	0.2231	0.10378	0.20577	0.31703	0.45816	0.16306	0.29083	0.39118	0.48564
	0.5108	0.20938	0.40583	0.59218	0.82804	0.28389	0.50146	0.65771	0.82653
	0.9163	0.32109	0.60238	0.83546	1.14197	0.39206	0.67338	0.86214	1.11994
	1.6094	0.46497	0.84849	1.12864	1.79687	0.48796	0.85048	1.11032	1.83578
	$n = 400$								

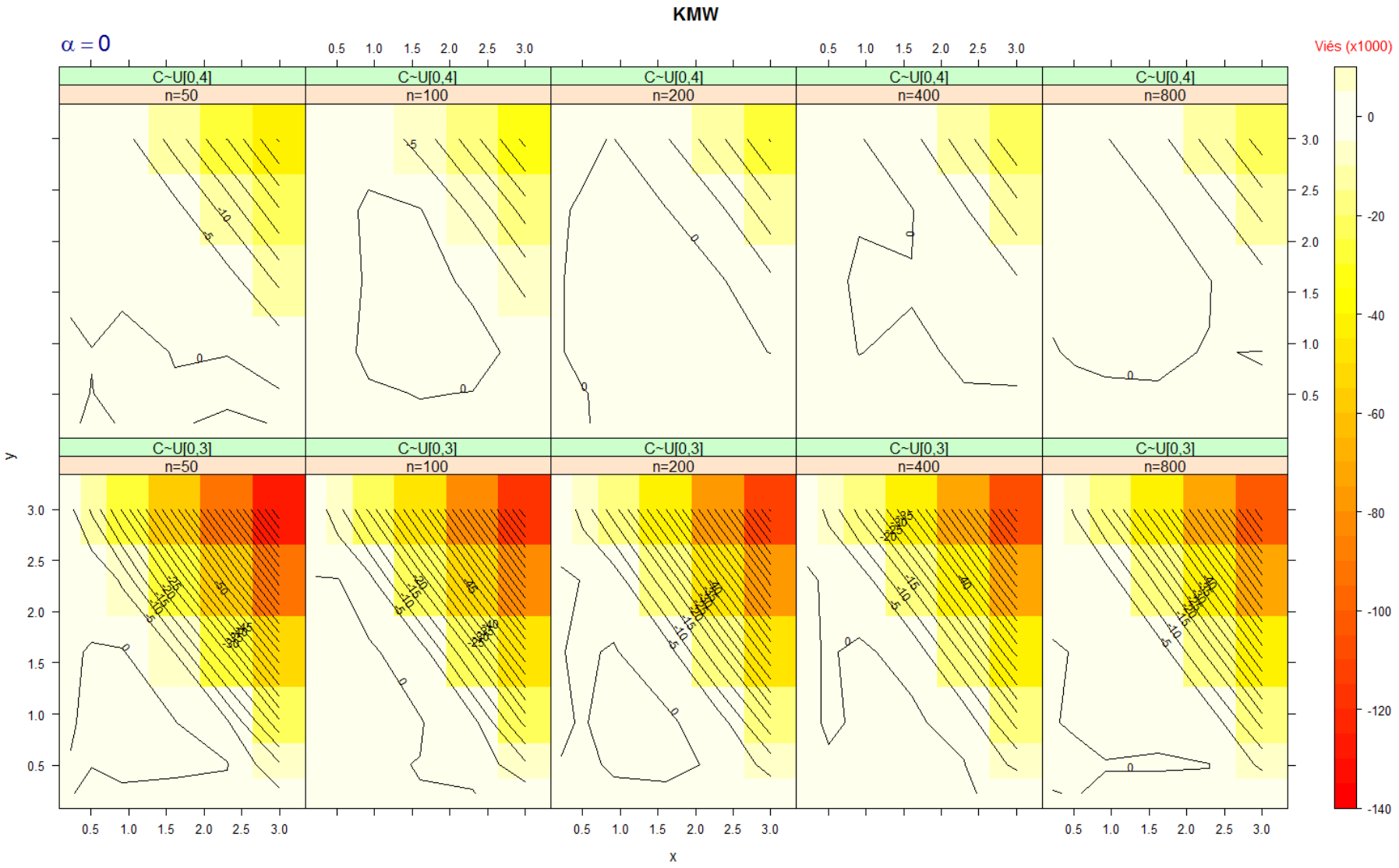


Figura B.1: Enviesamento de $\widehat{F}_{12}(x,y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 0$).

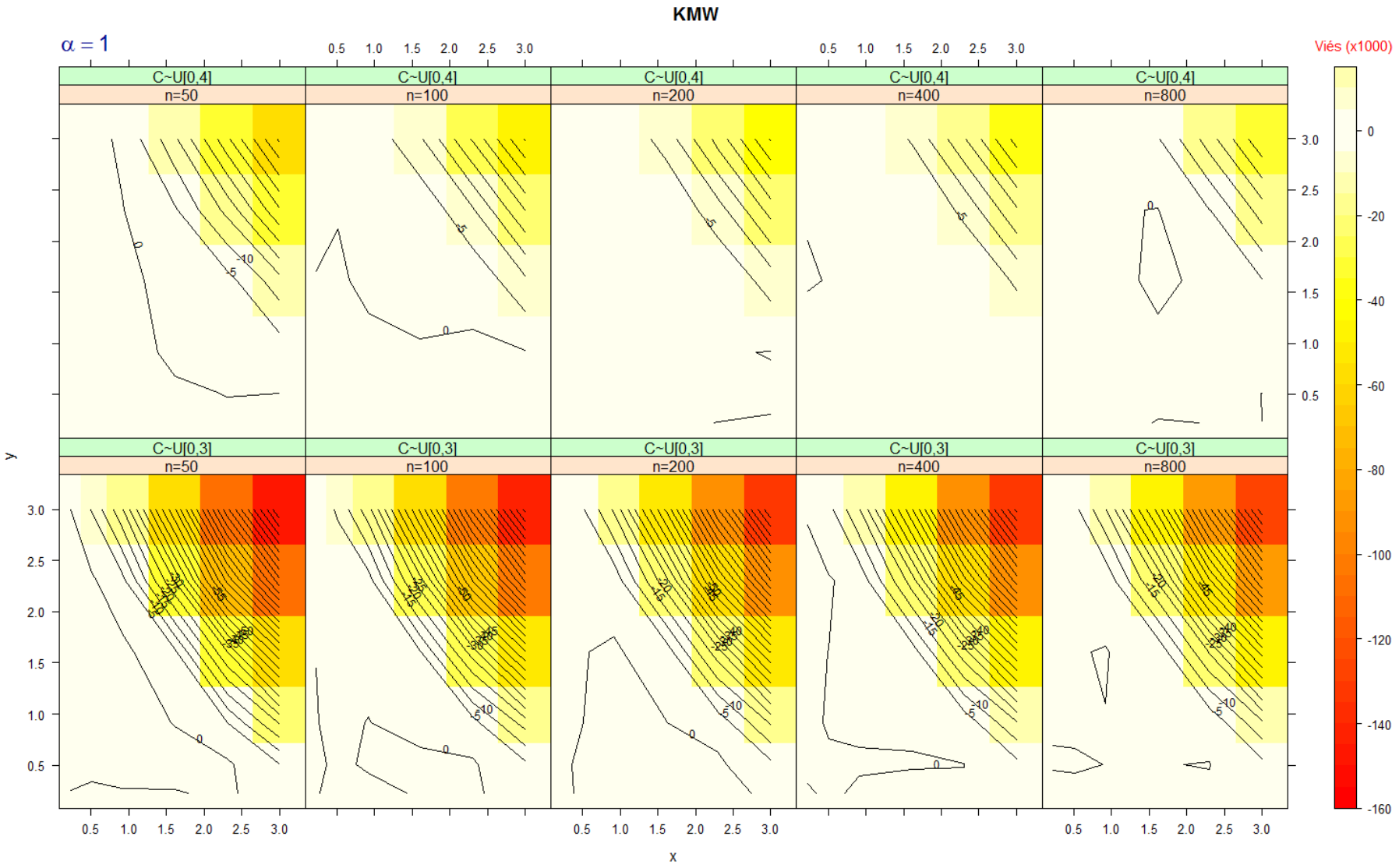


Figura B.2: Enviesamento de $\widehat{F}_{12}(x, y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 1$).

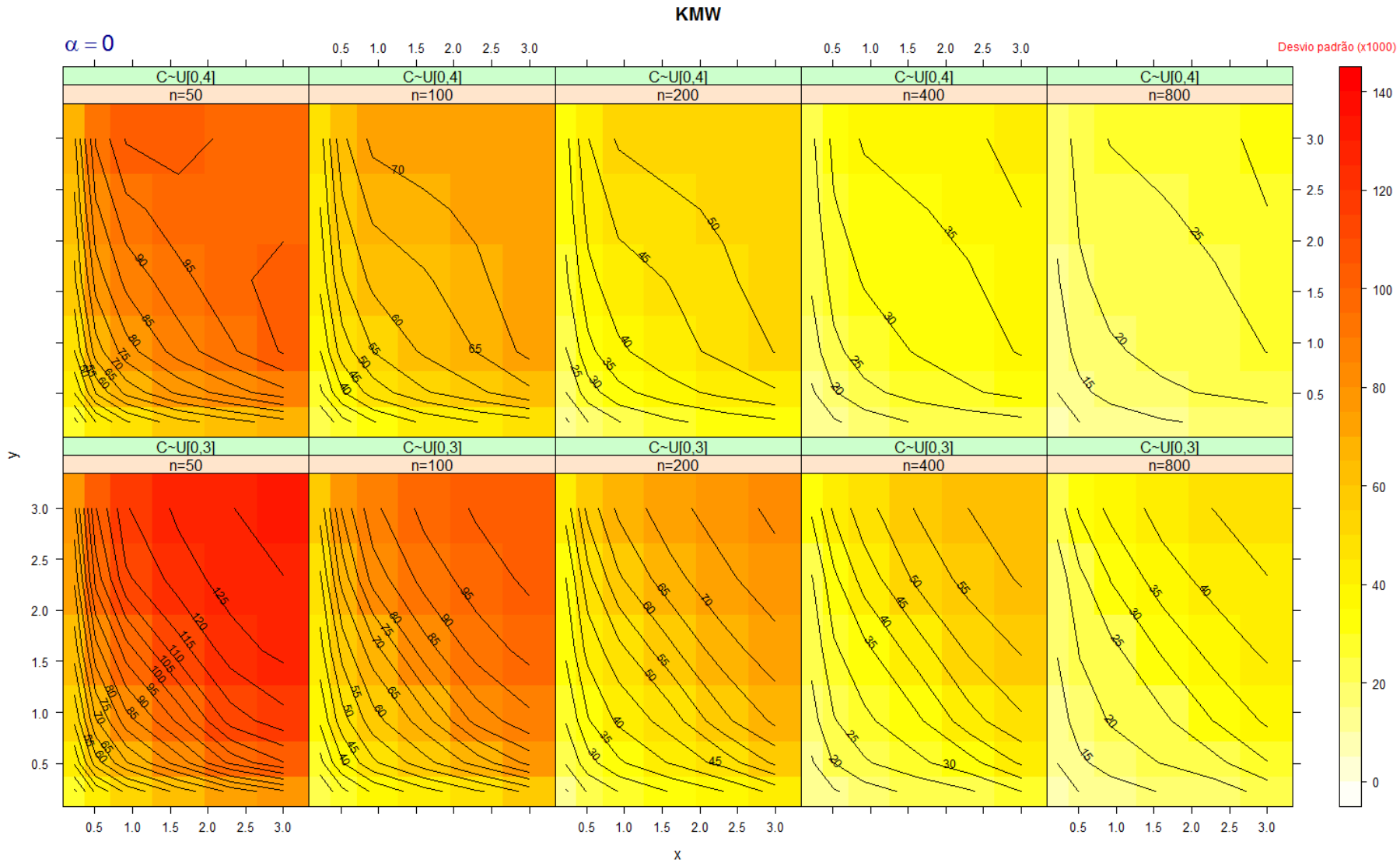


Figura B.3: Desvio padrão de $\widehat{F}_{12}(x,y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 0$).

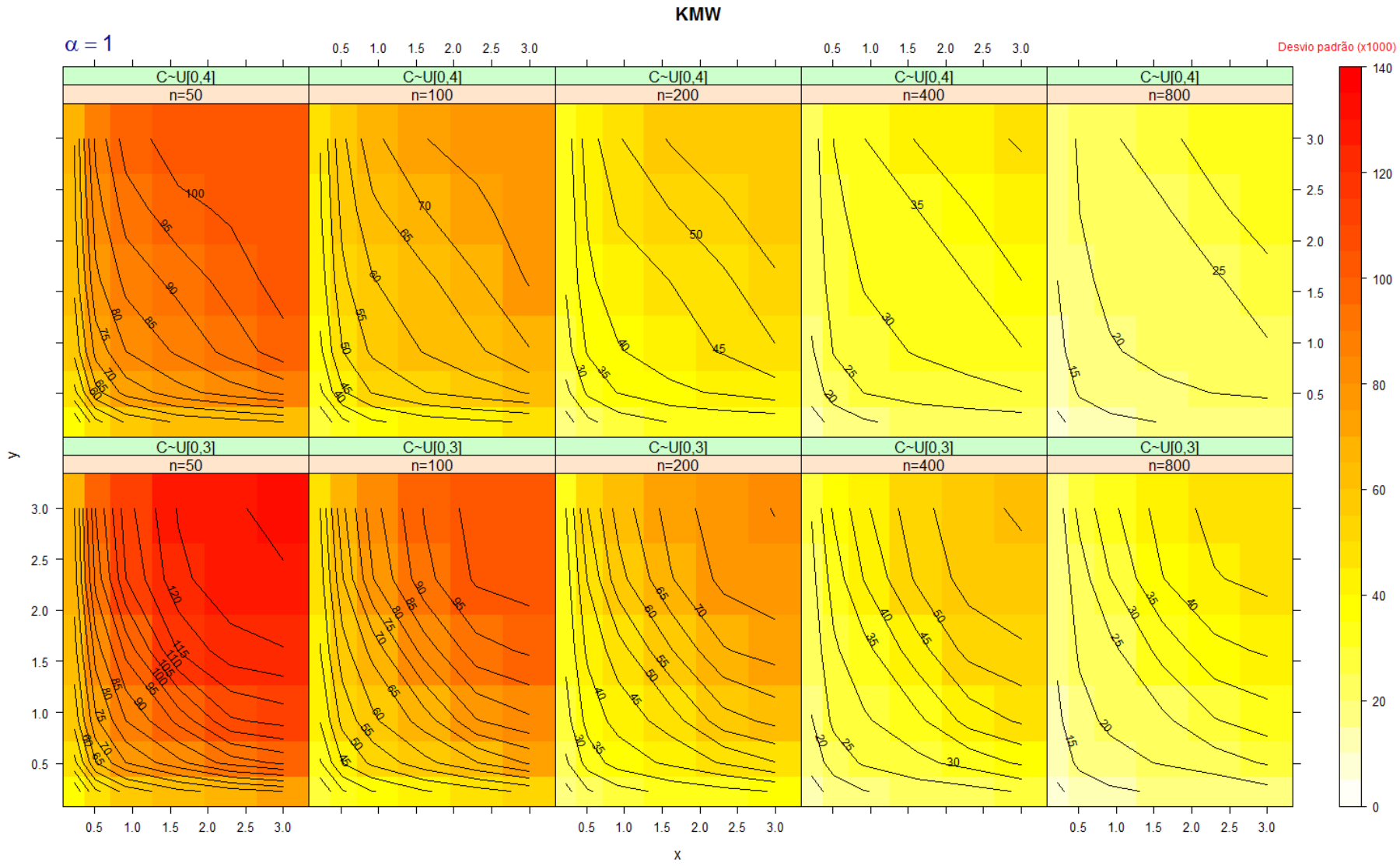


Figura B.4: Desvio padrão de $\widehat{F}_{12}(x, y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 1$).

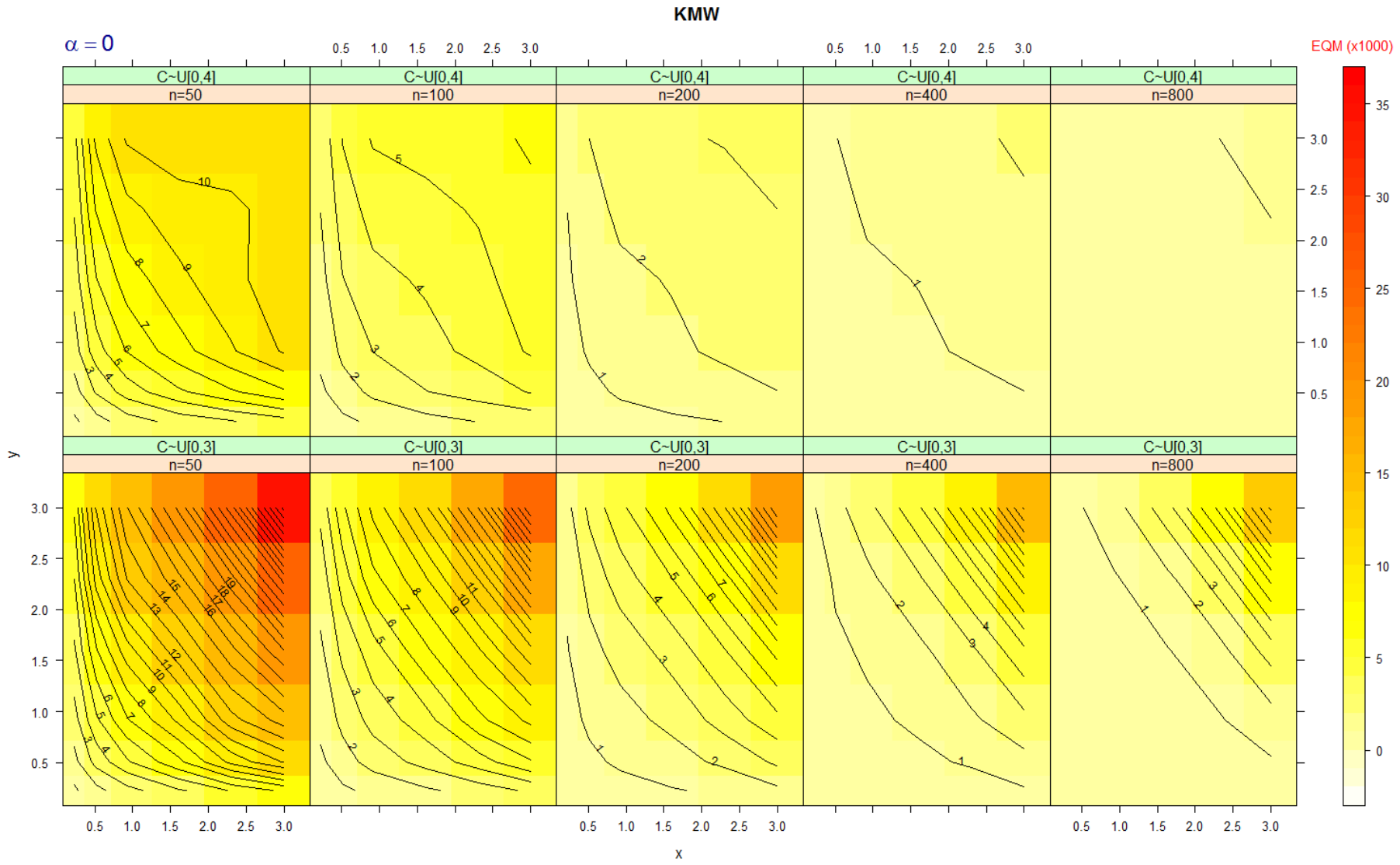


Figura B.5: Erro quadrático médio de $\widehat{F}_{12}(x,y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 0$).

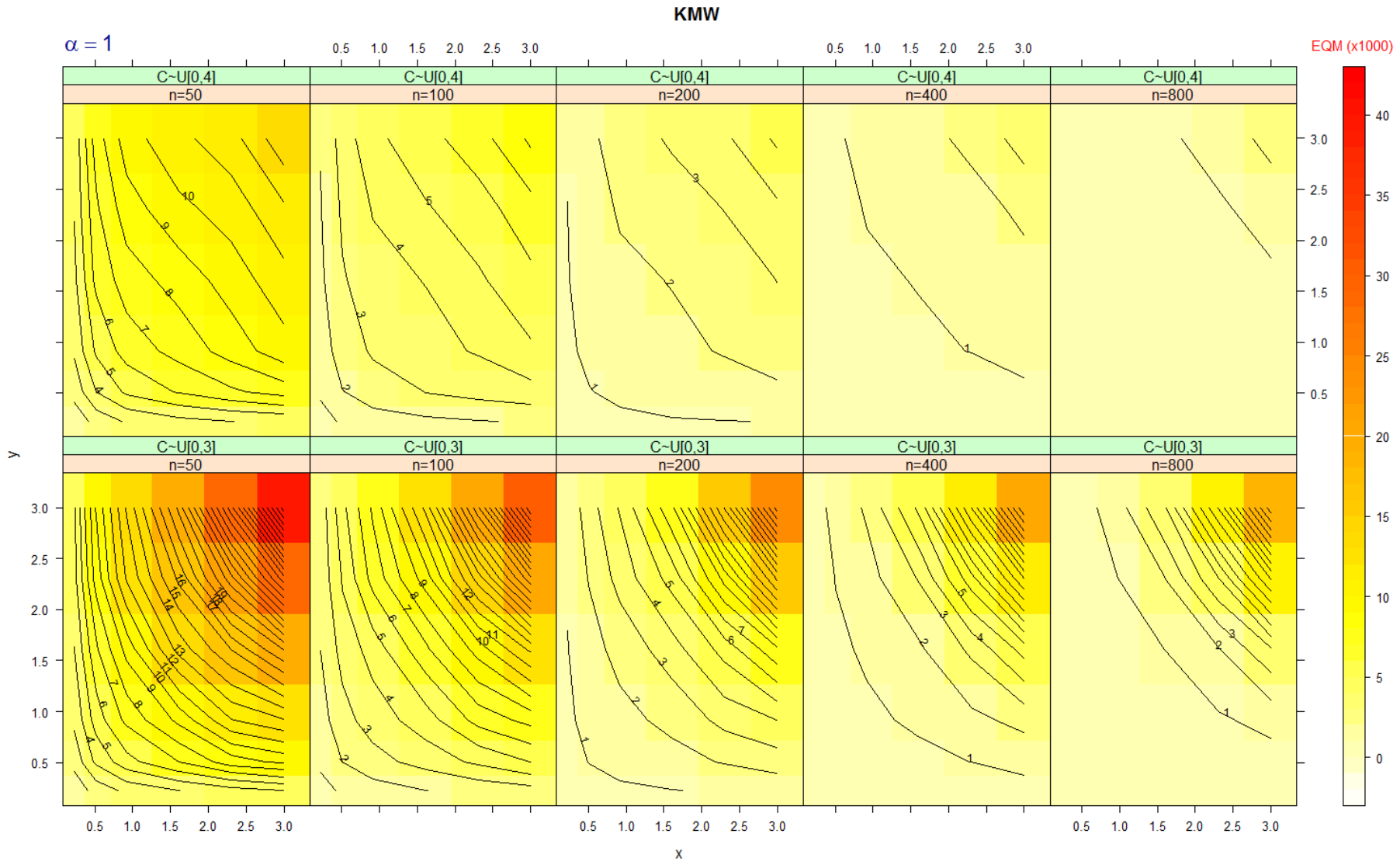


Figura B.6: Erro quadrático médio de $\widehat{F}_{12}(x,y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 1$).

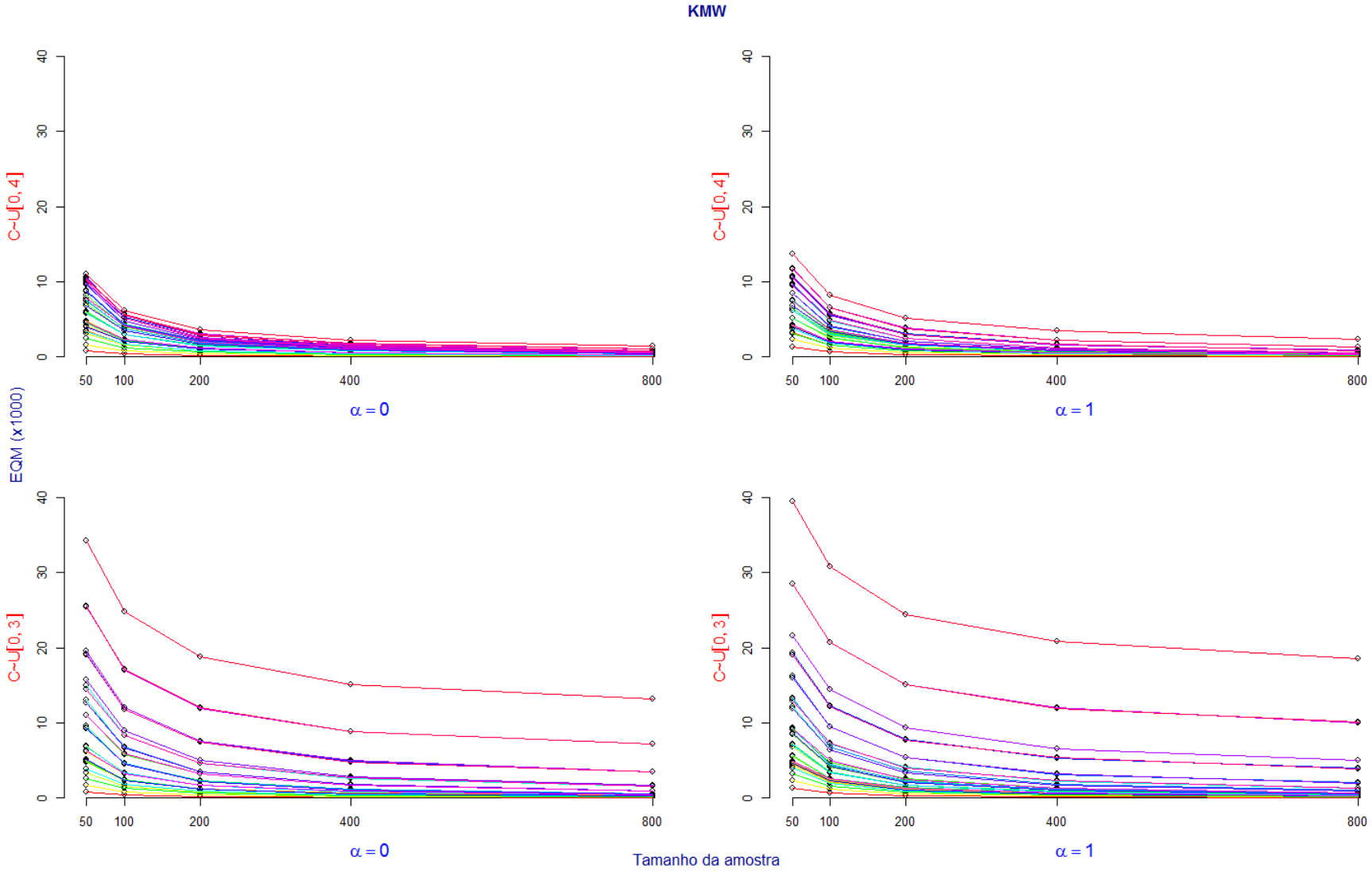


Figura B.7: Erro quadrático médio de $\widehat{F}_{12}(x, y)$ (x1000) versus o tamanho da amostra.

Anexo C Estimador Kaplan-Meier pesado pré-suavizado

Tabela C.1: Enviesamento de $\widehat{F}_{12}(x,y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita.

		$\alpha = 0$				$\alpha = 1$			
	Y	0.2231	0.5108	0.9163	1.6094	0.2231	0.5108	0.9163	1.6094
	x								
$C \sim U[0,4]$	0.2231	0.81245	1.39412	1.8664	0.95920	0.061376	0.16964	-0.49940	-1.63853
	0.5108	1.63341	2.56981	2.9444	0.79676	0.391642	0.89946	0.32013	-1.02302
	0.9163	1.89207	2.57221	2.0836	-1.53836	1.079961	2.32150	1.59444	-0.85465
	1.6094	0.26993	-0.52609	-2.5270	-7.49873	2.049718	2.49997	-0.42012	-7.17182
	$n = 50$								
	0.2231	0.926671	1.59950	1.6218	0.88649	-0.1503	-0.45883	-1.07624	-1.87598
	0.5108	1.858962	2.58877	2.6152	0.98750	0.3203	0.67230	0.22181	-0.79947
	0.9163	1.929305	2.58274	2.0055	-0.74791	1.2980	2.18354	1.53634	-0.31054
	1.6094	0.947235	0.31101	-1.4803	-5.26659	2.4844	2.95049	-0.19449	-5.79327
	$n = 100$								
	0.2231	1.20060	2.1118740	2.48371	1.67760	-0.1503	-0.45883	-1.07624	-1.87598
	0.5108	2.06134	3.4382510	3.58560	1.48815	0.3203	0.67230	0.22181	-0.79947
	0.9163	2.31951	3.5118600	2.62014	-0.99239	1.2980	2.18354	1.53634	-0.31054
	1.6094	1.32110	1.4319145	-0.82504	-5.24422	2.4844	2.95049	-0.19449	-5.79327
	$n = 200$								
	0.2231	1.4782	2.45446	2.66815	1.7571	-0.37656	-1.02865	-1.63428	-2.50453
	0.5108	2.5695	3.70406	3.49139	1.3563	0.10523	-0.21594	-0.93550	-1.93288
	0.9163	2.7588	3.57359	2.23704	-1.2114	1.32814	1.59840	0.80744	-0.84207
	1.6094	1.9742	1.74426	-1.05121	-5.3096	3.16451	3.50042	0.60325	-4.51647
	$n = 400$								
$C \sim U[0,3]$	0.2231	1.72614	2.5221	2.42770	0.1181	0.5549	0.32623	-0.28384	-1.5914
	0.5108	2.59381	4.2034	3.14540	-1.3765	2.1155	2.55306	1.96517	-0.3301
	0.9163	2.37182	3.1294	-0.05539	-7.1429	2.7666	2.74164	0.19439	-4.1302
	1.6094	0.10819	-1.9453	-7.37962	-14.8240	2.7126	-0.04872	-7.32433	-15.7234
	$n = 50$								
	0.2231	1.91017	3.00958	3.09194	0.92193	0.84681	0.43718	-0.20426	-1.34866
	0.5108	3.06162	4.38062	3.60360	-0.86427	1.84064	1.93586	0.86148	-0.81101
	0.9163	3.09130	3.17825	0.27533	-5.98611	2.60471	2.44659	-0.28414	-3.91423
	1.6094	1.10076	-1.10682	-5.51476	-7.56587	2.78065	0.12050	-7.31761	-11.36238
	$n = 100$								
	0.2231	1.9365	3.08027	2.89618	0.99794	0.70784	0.80327	0.18573	-0.91622
	0.5108	3.1759	4.32890	3.19036	-0.93436	1.67886	2.10338	1.19923	-0.62859
	0.9163	3.1430	3.25400	0.00611	-5.39680	2.20299	2.58785	0.31642	-3.09462
	1.6094	1.2284	-1.00970	-5.79196	-3.80376	2.66978	0.94229	-5.74208	-6.31673
	$n = 200$								
	0.2231	2.27703	3.40264	3.38791	1.38648	0.81741	0.95743	-0.087686	-1.4172
	0.5108	3.32898	4.73153	3.57032	-0.58471	1.87867	2.11614	0.744062	-1.2369
	0.9163	3.27538	3.32562	0.25472	-5.22030	2.76315	2.68490	-0.085354	-3.2525
	1.6094	1.32553	-0.82931	-5.22608	-0.57187	3.48268	1.50529	-5.515983	-3.0217
	$n = 400$								

Tabela C.2: Desvio padrão de $\widehat{F}_{12}(x,y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita.

		$\alpha = 0$				$\alpha = 1$			
	Y	0.2231	0.5108	0.9163	1.6094	0.2231	0.5108	0.9163	1.6094
	x								
$C \sim U[0,4]$	0.2231	26.475	37.460	45.977	53.137	34.524	46.461	53.394	57.892
	0.5108	37.902	52.842	63.431	71.786	46.169	60.833	69.429	74.669
	0.9163	46.196	63.697	73.902	82.079	53.127	69.307	76.996	81.551
	1.6094	53.729	72.317	81.455	86.319	57.164	74.654	81.427	83.575
	$n = 50$								
	0.2231	18.577	26.074	32.018	37.325	24.199	32.095	37.010	40.137
	0.5108	26.920	37.214	44.532	50.771	32.199	42.342	48.368	52.089
	0.9163	32.850	44.865	52.035	57.125	37.032	47.699	53.503	56.923
	1.6094	37.933	50.993	57.242	59.604	39.887	50.908	55.983	58.026
	$n = 100$								
	0.2231	13.256	18.555	23.014	26.421	24.199	32.095	37.010	40.137
	0.5108	18.657	25.994	31.500	35.658	32.199	42.342	48.368	52.089
	0.9163	23.141	31.541	36.886	40.287	37.032	47.699	53.503	56.923
	1.6094	26.645	36.016	40.314	41.684	39.887	50.908	55.983	58.026
	$n = 200$								
	0.2231	9.3666	13.359	16.241	18.848	12.091	16.157	18.567	20.136
	0.5108	13.2315	18.421	22.054	25.235	16.102	21.336	24.290	26.140
	0.9163	16.4743	22.397	25.922	28.480	18.499	24.282	26.997	28.719
	1.6094	18.8361	25.212	28.357	29.432	19.936	25.972	28.478	29.221
	$n = 400$								
$C \sim U[0,3]$	0.2231	26.690	38.034	46.885	55.858	34.209	46.272	53.770	59.122
	0.5108	37.852	53.913	64.811	76.290	45.663	61.698	71.256	79.158
	0.9163	46.646	64.617	76.468	88.503	52.961	70.468	79.686	89.263
	1.6094	55.589	76.422	89.391	99.992	57.757	76.761	87.796	100.885
	$n = 50$								
	0.2231	18.605	26.326	32.498	38.318	24.262	32.223	37.364	41.220
	0.5108	25.959	36.827	45.074	53.255	32.265	42.727	48.818	55.104
	0.9163	32.329	44.335	53.222	61.563	37.379	48.655	55.295	62.388
	1.6094	38.197	52.094	61.495	71.237	40.849	53.607	61.193	71.220
	$n = 100$								
	0.2231	13.296	18.658	22.878	26.868	17.307	22.941	26.475	28.947
	0.5108	18.765	26.062	31.583	37.148	22.894	30.560	34.905	38.630
	0.9163	22.987	31.917	37.629	43.210	26.263	34.686	39.270	43.678
	1.6094	27.282	37.658	43.840	51.731	28.519	37.950	43.258	51.085
	$n = 200$								
	0.2231	9.324	13.116	16.142	19.017	12.053	16.085	18.547	20.432
	0.5108	13.248	18.576	22.440	26.197	15.909	21.318	24.414	27.108
	0.9163	16.285	22.658	26.590	30.636	18.397	24.426	27.506	30.801
	1.6094	19.009	26.473	30.419	37.292	20.058	26.848	30.250	37.111
	$n = 400$								

Tabela C.3: Erro quadrático médio de $\widehat{F}_{12}(x,y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita.

		$\alpha = 0$				$\alpha = 1$			
	Y	0.2231	0.5108	0.9163	1.6094	0.2231	0.5108	0.9163	1.6094
	x								
$C \sim U[0,4]$	0.2231	0.70151	1.4051	2.1172	2.8242	1.1918	2.1584	2.8508	3.3538
	0.5108	1.43905	2.7986	4.0317	5.1533	2.1315	3.7011	4.8201	5.5760
	0.9163	2.13744	4.0635	5.4653	6.7387	2.8234	4.8084	5.9304	6.6507
	1.6094	2.88655	5.2294	6.6407	7.5064	3.2716	5.5789	6.6298	7.0355
	$n = 50$								
	0.2231	0.34592	0.68233	1.0277	1.3938	0.58557	1.0302	1.3707	1.6143
	0.5108	0.72807	1.39143	1.9898	2.5784	1.03679	1.7931	2.3393	2.7136
	0.9163	1.08273	2.01937	2.7113	3.2635	1.37290	2.2797	2.8646	3.2399
	1.6094	1.43968	2.60008	3.2786	3.5801	1.59696	2.6000	3.1338	3.4002
	$n = 100$								
	0.2231	0.17715	0.34870	0.53578	0.70082	0.58557	1.0302	1.3707	1.6143
	0.5108	0.35228	0.68744	1.00499	1.27356	1.03679	1.7931	2.3393	2.7136
	0.9163	0.54081	1.00704	1.36727	1.62388	1.37290	2.2797	2.8646	3.2399
	1.6094	0.71163	1.29911	1.62570	1.76490	1.59696	2.6000	3.1338	3.4002
	$n = 200$								
	0.2231	0.089909	0.18448	0.27086	0.35828	0.14632	0.26210	0.34736	0.41169
	0.5108	0.181657	0.35302	0.49852	0.63857	0.25927	0.45521	0.59084	0.68695
	0.9163	0.278985	0.51435	0.67691	0.81250	0.34395	0.59213	0.72941	0.82540
	1.6094	0.358661	0.63862	0.80512	0.89438	0.40741	0.68675	0.81126	0.87417
	$n = 400$								
$C \sim U[0,3]$	0.2231	0.71525	1.4528	2.2039	3.1198	1.1704	2.1410	2.8910	3.4976
	0.5108	1.43934	2.9240	4.2099	5.8214	2.0894	3.8128	5.0808	6.2654
	0.9163	2.18125	4.1847	5.8468	7.8830	2.8122	4.9728	6.3493	7.9842
	1.6094	3.08981	5.8435	8.0444	10.2172	3.3429	5.8917	7.7609	10.4240
	$n = 50$								
	0.2231	0.34977	0.70204	1.0656	1.4690	0.58933	1.0384	1.3960	1.7008
	0.5108	0.68316	1.37528	2.0445	2.8366	1.04432	1.8291	2.3837	3.0368
	0.9163	1.05460	1.97548	2.8324	3.8254	1.40385	2.3730	3.0574	3.9072
	1.6094	1.46005	2.71469	3.8117	5.1315	1.67618	2.8735	3.7978	5.2009
	$n = 100$								
	0.2231	0.18053	0.35759	0.53175	0.72282	0.30000	0.52689	0.7009	0.83866
	0.5108	0.36216	0.69791	1.00759	1.38074	0.52692	0.93825	1.2197	1.49254
	0.9163	0.53824	1.02917	1.41577	1.89608	0.69454	1.20971	1.5421	1.91718
	1.6094	0.74575	1.41898	1.95533	2.69025	0.82038	1.44098	1.9040	2.64934
	$n = 200$								
	0.2231	0.092113	0.18360	0.27203	0.36355	0.14594	0.25963	0.34395	0.41944
	0.5108	0.186579	0.36743	0.51625	0.68654	0.25661	0.45891	0.59652	0.73629
	0.9163	0.275917	0.52439	0.70704	0.96570	0.34604	0.60380	0.75651	0.95917
	1.6094	0.363061	0.70144	0.95251	1.39090	0.41442	0.72300	0.94543	1.38619
	$n = 400$								

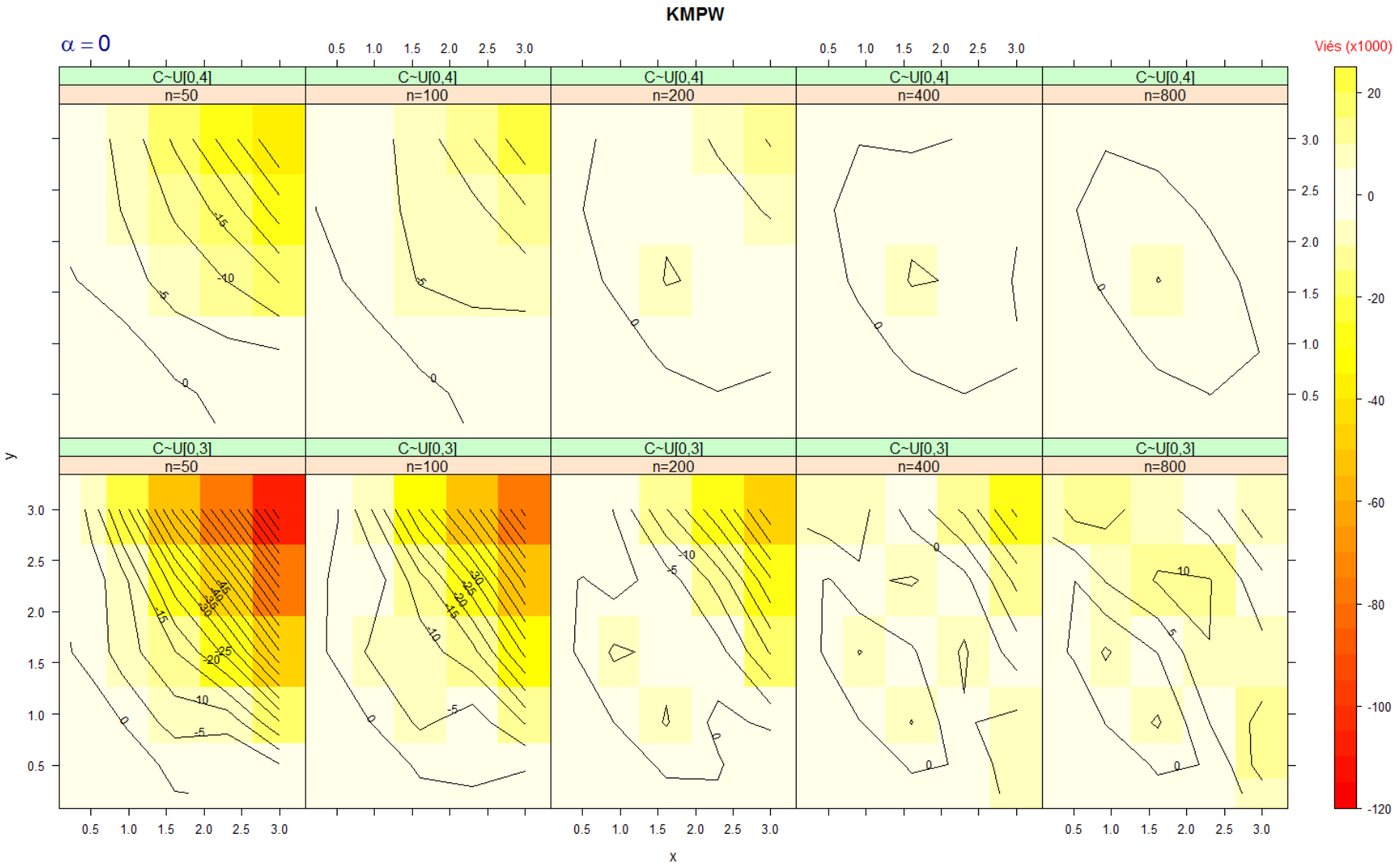


Figura C.1: Enviesamento de $\widehat{F}_{12}(x,y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 0$).

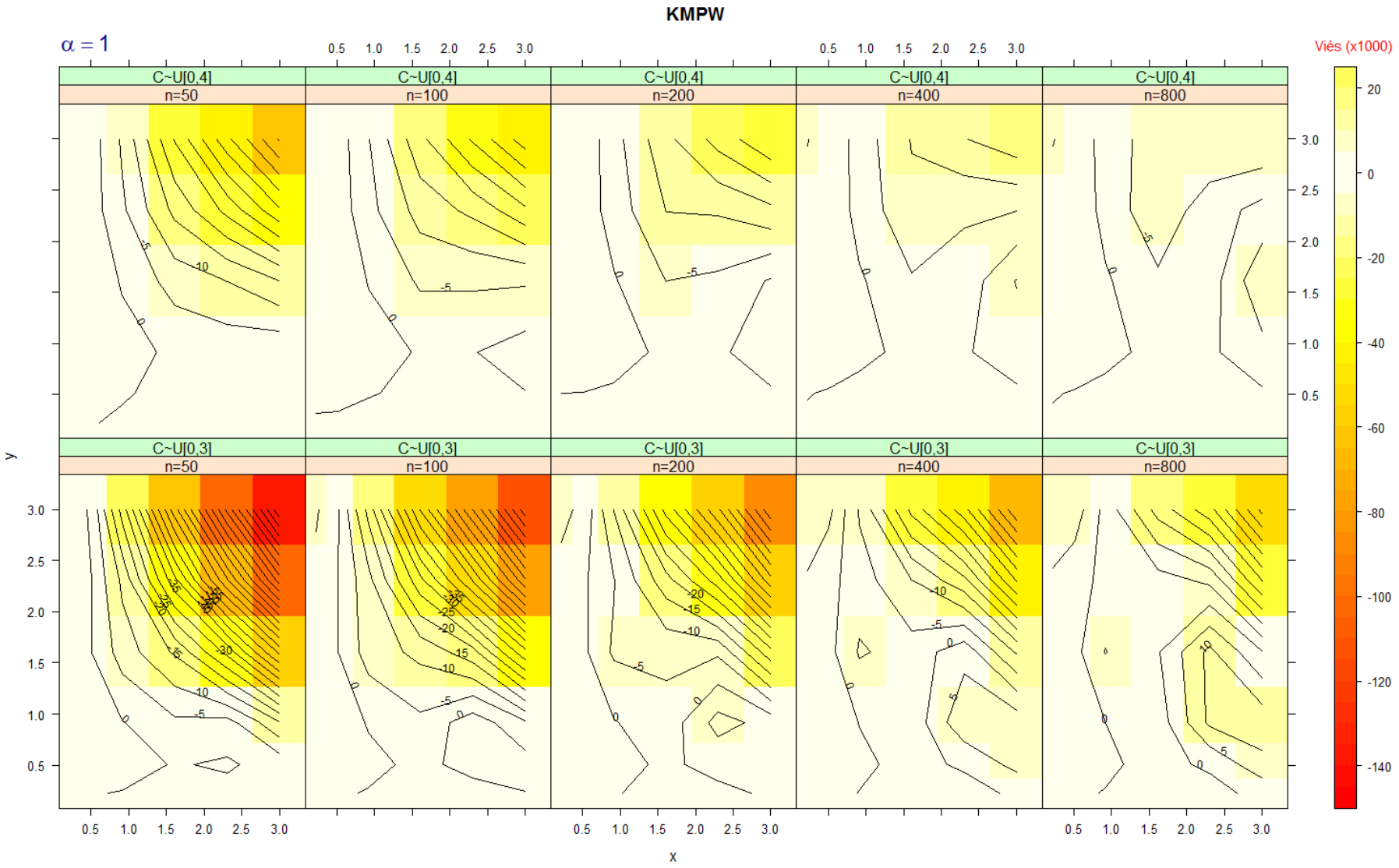


Figura C.2: Enviesamento de $\widehat{F}_{12}(x,y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 1$).

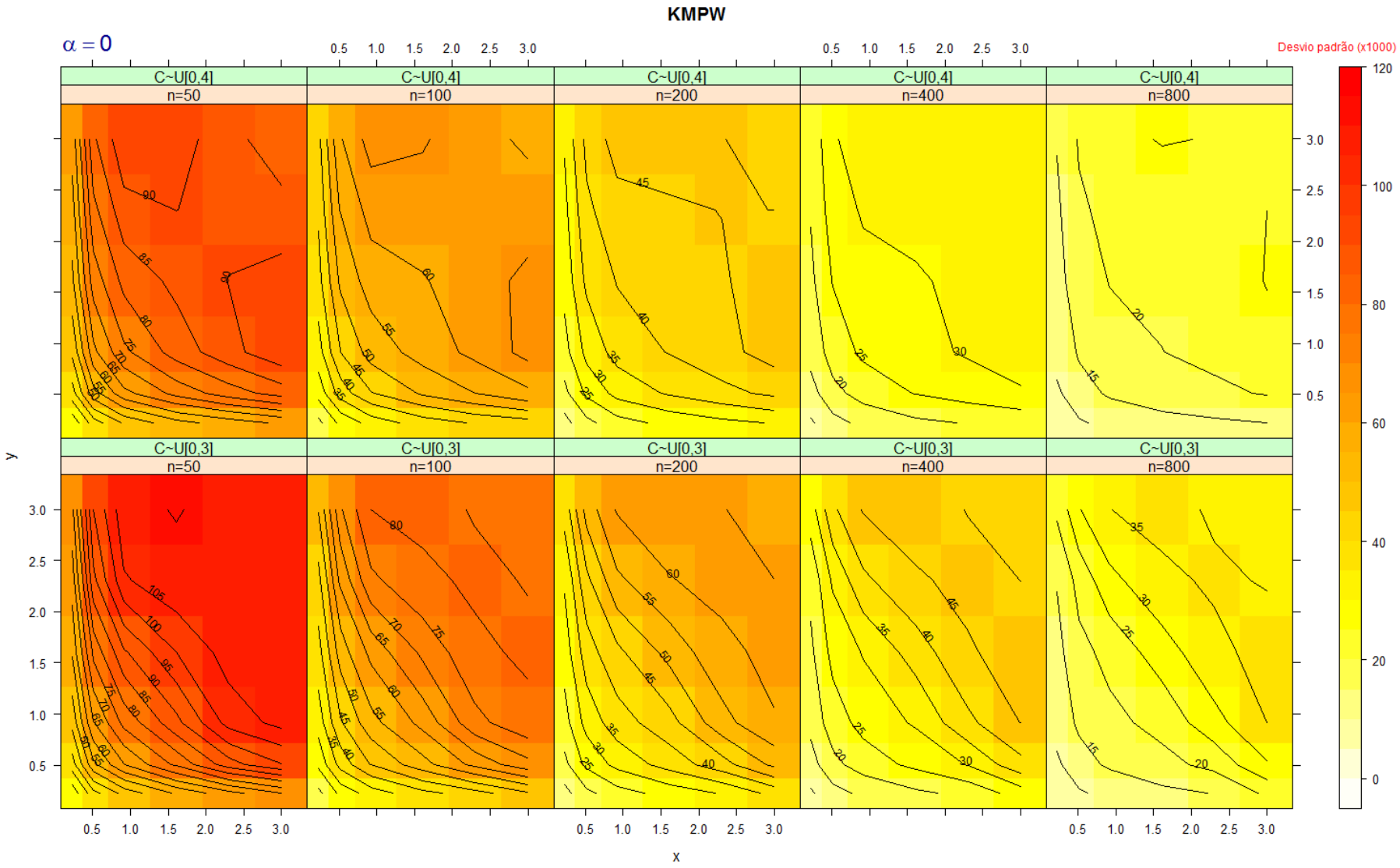


Figura C.3: Desvio padrão de $\widehat{F}_{12}(x,y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 0$).

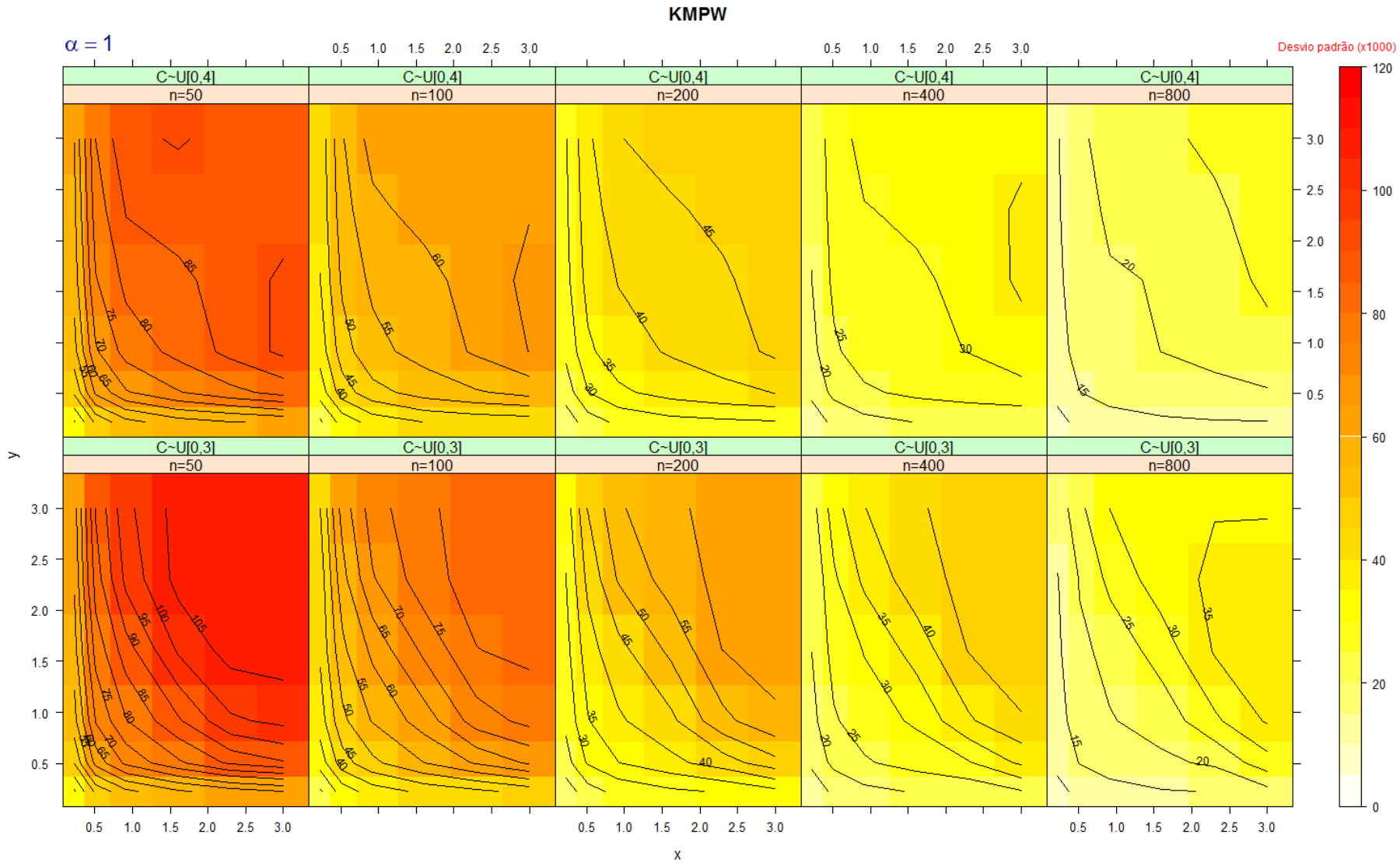


Figura C.4: Desvio padrão de $\widehat{F}_{12}(x,y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 1$).

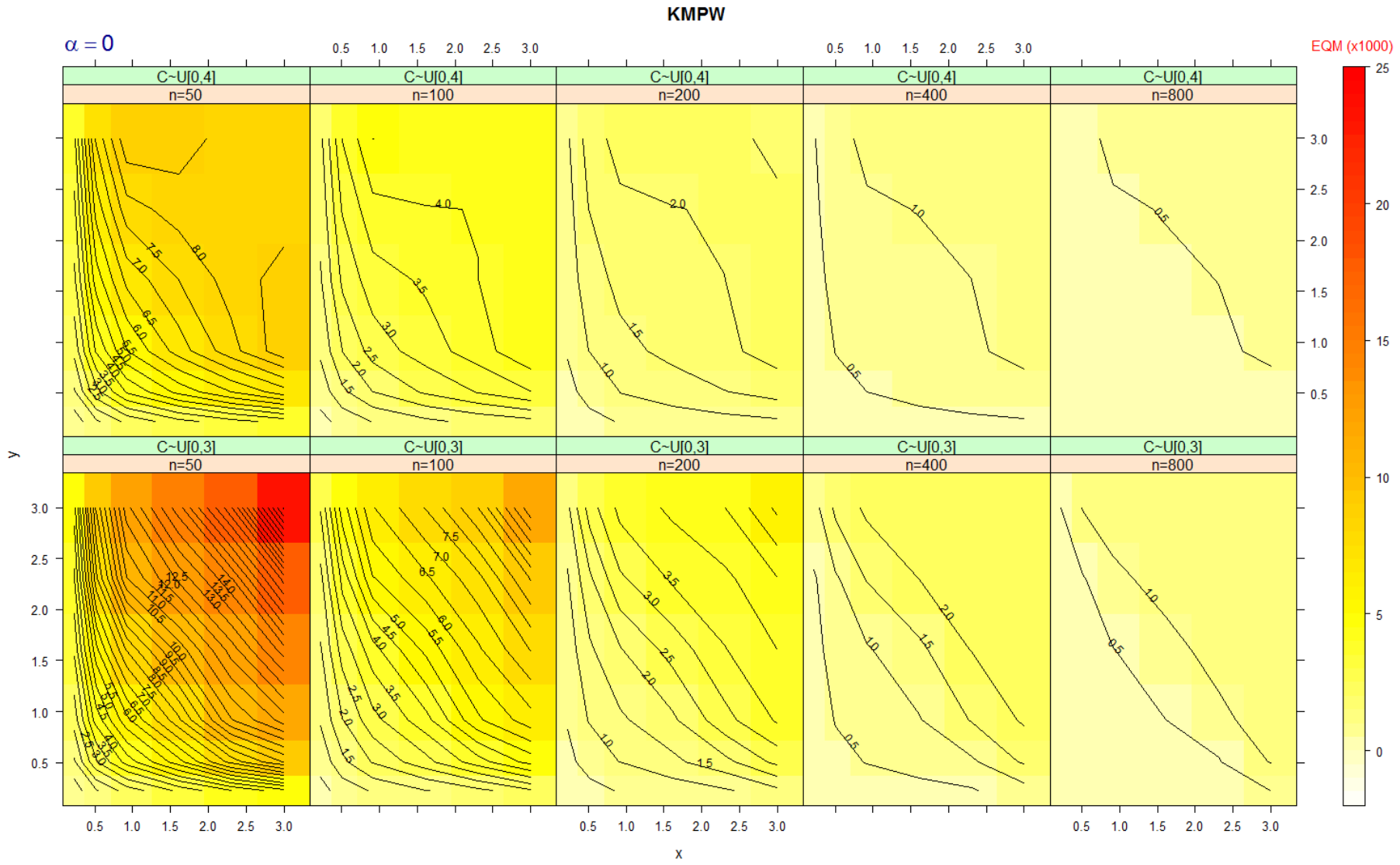


Figura C.5: Erro quadrático médio de $\widehat{F}_{12}(x, y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 0$).

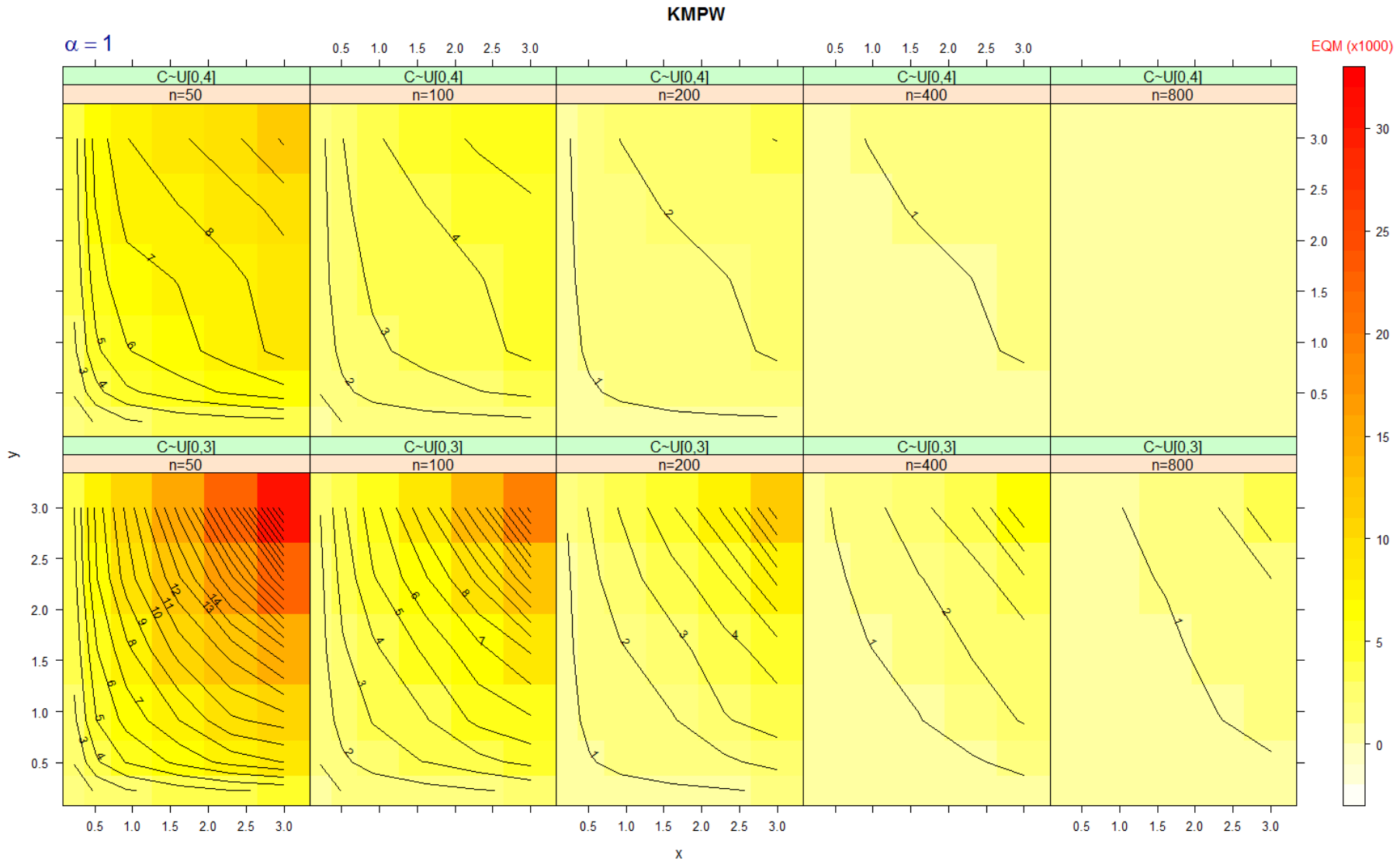


Figura C.6: Erro quadrático médio de $\widehat{F}_{12}(x, y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 1$).

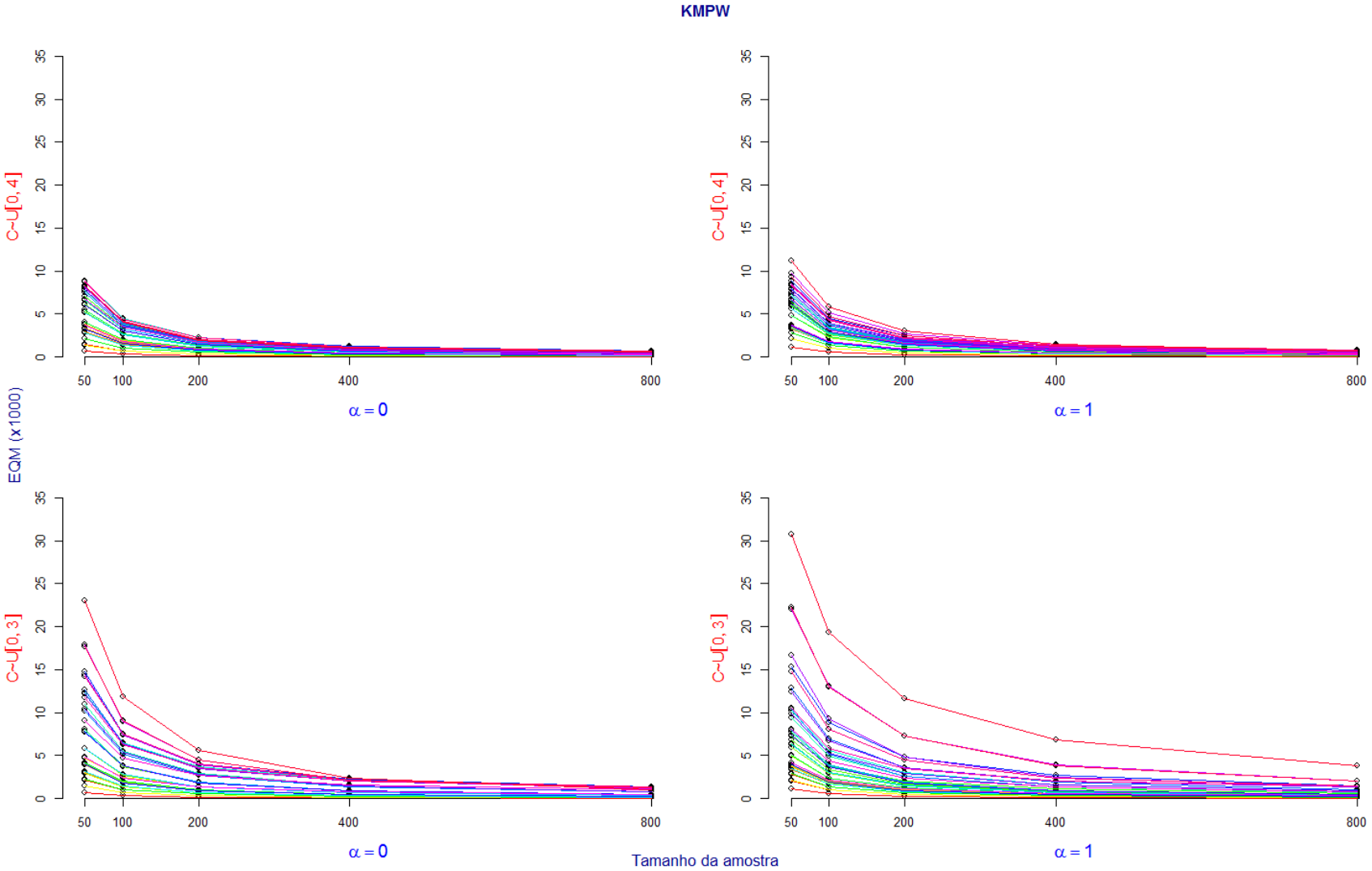


Figura C.7: Erro quadrático médio de $\widehat{F}_{12}(x, y)$ (x1000) versus o tamanho da amostra.

Anexo D Estimador de Lin

Tabela D.1: Enviesamento de $\widehat{F}_{12}(x,y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita.

		$\alpha = 0$				$\alpha = 1$			
	Y	0.2231	0.5108	0.9163	1.6094	0.2231	0.5108	0.9163	1.6094
	x								
$C \sim U[0,4]$	0.2231	0.73645	0.062799	0.39292	0.59143	0.67562	1.04220	0.99146	0.535469
	0.5108	0.31550	-0.071191	0.44513	0.64845	0.44957	0.91829	0.70373	0.223551
	0.9163	0.43142	0.055604	0.43923	0.39407	0.33655	0.96702	0.59547	-0.228517
	1.6094	-0.38248	-0.869972	-0.48829	-0.33988	0.20967	0.29528	0.30968	-0.012111
	$n = 50$								
	0.2231	-0.17526	-0.42136	-0.562334	-0.14104	0.110568	0.157329	0.067191	0.045232
	0.5108	-0.14347	-0.68666	-0.541024	0.25056	0.128263	0.532704	0.535970	0.449718
	0.9163	-0.30596	-0.62588	-0.295933	0.77937	0.176298	0.559425	0.424627	0.142135
	1.6094	-0.36793	-0.91766	-0.367439	0.70112	0.085889	0.403753	-0.089779	-0.285637
	$n = 100$								
	0.2231	-0.173151	-0.143243	0.10250	0.13739	0.110568	0.157329	0.067191	0.045232
	0.5108	-0.176783	-0.017566	0.37611	0.23759	0.128263	0.532704	0.535970	0.449718
	0.9163	-0.043538	0.332612	0.53818	0.11858	0.176298	0.559425	0.424627	0.142135
	1.6094	-0.203297	0.253152	0.49223	0.05498	0.085889	0.403753	-0.089779	-0.285637
	$n = 200$								
	0.2231	0.017974	0.11451	0.16490	0.0623573	-0.21415	-0.40575	-0.30121	-0.39735
	0.5108	0.282864	0.14876	0.10358	0.0312666	-0.15389	-0.25913	-0.33095	-0.31019
	0.9163	0.282431	0.16949	-0.12801	0.0018966	-0.05197	-0.20406	-0.32551	-0.09369
	1.6094	0.356678	0.46489	0.13150	0.2690721	0.05627	0.02285	-0.20878	0.21291
	$n = 400$								
$C \sim U[0,3]$	0.2231	0.968578	-0.048996	0.014936	-0.17789	0.34342	-0.24335	-0.162968	-0.24440
	0.5108	0.074431	0.144850	0.393099	0.58550	1.00795	0.81032	1.282371	1.14662
	0.9163	-0.144944	0.564509	0.858473	1.02590	1.12384	0.97729	0.926582	0.87784
	1.6094	-0.121372	0.279669	1.185577	19.77474	1.33984	0.13173	0.083613	27.36708
	$n = 50$								
	0.2231	0.150369	-0.032354	0.46563	0.26471	0.16617	-0.32836	-0.17885	-0.20260
	0.5108	0.090099	0.038448	0.87874	0.91675	0.39264	0.14744	0.14002	0.25381
	0.9163	0.384591	0.083910	0.80107	0.66771	0.57285	0.52514	0.49330	-0.64221
	1.6094	0.099861	-0.245924	0.83862	16.23969	0.56064	0.35755	0.36139	20.85249
	$n = 100$								
	0.2231	-0.120756	-0.124776	-0.10112	-0.10354	-0.21853	-0.09271	0.11585	0.293950
	0.5108	-0.028546	-0.255819	0.08584	0.18097	-0.13516	0.01450	0.34982	0.529428
	0.9163	0.184438	0.033432	0.18706	0.90222	-0.51082	-0.15569	0.16767	0.074126
	1.6094	0.194810	-0.113094	0.27190	13.44749	-0.55267	-0.24118	0.29081	17.732328
	$n = 200$								
	0.2231	0.210543	0.161498	0.37761	0.099859	-0.02635	0.22519	0.03093	0.003535
	0.5108	0.092153	0.183715	0.40166	0.282220	0.10414	0.10817	0.04754	0.019343
	0.9163	0.247823	-0.051353	0.35053	0.703197	0.11355	-0.11412	-0.28945	-0.098907
	1.6094	0.139373	-0.247965	0.42756	12.335085	0.13495	-0.11963	-0.66876	15.426430
	$n = 400$								

Tabela D.2: Desvio padrão de $\widehat{F}_{12}(x,y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita.

		$\alpha = 0$				$\alpha = 1$			
	Y	0.2231	0.5108	0.9163	1.6094	0.2231	0.5108	0.9163	1.6094
	x								
$C \sim U[0,4]$	0.2231	29.799	42.292	51.243	57.392	37.672	49.782	55.817	58.733
	0.5108	43.272	59.479	70.302	76.880	50.671	65.064	72.718	75.271
	0.9163	53.172	71.615	82.199	89.047	59.034	75.195	82.833	85.299
	1.6094	64.184	82.621	92.628	99.123	66.234	84.259	92.763	98.027
	$n = 50$								
	0.2231	21.612	29.909	35.925	40.389	26.615	34.405	38.540	40.488
	0.5108	31.071	42.178	49.439	54.300	35.409	45.356	50.530	52.614
	0.9163	37.896	50.563	57.720	61.823	41.252	51.793	57.329	59.360
	1.6094	45.532	58.173	64.816	67.822	46.609	57.503	63.616	66.853
	$n = 100$								
	0.2231	15.515	21.373	25.955	28.704	26.615	34.405	38.540	40.488
	0.5108	21.591	29.536	34.938	38.368	35.409	45.356	50.530	52.614
	0.9163	26.886	35.778	41.179	43.857	41.252	51.793	57.329	59.360
	1.6094	31.806	41.220	45.600	47.545	46.609	57.503	63.616	66.853
	$n = 200$								
	0.2231	10.901	15.272	18.140	20.451	13.300	17.399	19.493	20.454
	0.5108	15.299	20.776	24.355	27.001	17.767	22.993	25.512	26.590
	0.9163	19.150	25.283	28.758	30.897	20.633	26.563	29.005	30.129
	1.6094	22.563	28.780	31.867	33.182	23.242	29.414	32.272	33.506
	$n = 400$								
$C \sim U[0,3]$	0.2231	30.791	44.137	53.637	62.246	37.939	50.322	56.681	59.793
	0.5108	44.809	62.957	74.145	85.575	51.433	67.556	76.016	81.791
	0.9163	56.123	75.653	89.523	105.247	61.075	79.428	89.874	101.600
	1.6094	71.228	93.162	111.543	121.642	72.738	92.634	111.896	123.211
	$n = 50$								
	0.2231	22.412	31.454	37.936	43.590	27.117	35.162	39.623	41.963
	0.5108	31.366	43.349	51.557	59.560	36.520	46.950	52.147	56.701
	0.9163	39.617	52.652	62.373	72.892	43.215	54.882	62.094	70.818
	1.6094	49.345	63.922	76.113	88.821	51.361	65.190	78.053	90.927
	$n = 100$								
	0.2231	16.215	22.451	26.907	30.602	19.441	25.122	28.082	29.626
	0.5108	22.651	30.957	36.764	42.178	25.868	33.423	37.262	39.685
	0.9163	27.960	38.046	44.446	51.246	30.403	39.158	44.045	48.703
	1.6094	35.080	45.739	54.046	67.002	35.545	45.722	54.169	69.784
	$n = 200$								
	0.2231	11.336	15.618	18.779	21.493	13.550	17.631	19.707	20.969
	0.5108	15.985	21.905	25.763	29.151	17.949	23.501	26.296	28.162
	0.9163	19.821	26.739	31.235	35.956	21.260	27.561	31.088	34.475
	1.6094	24.518	32.361	37.361	50.368	25.015	32.219	37.913	52.387
	$n = 400$								

Tabela D.3: Erro quadrático médio de $\widehat{F}_{12}(x,y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita.

		$\alpha = 0$				$\alpha = 1$			
	Y	0.2231	0.5108	0.9163	1.6094	0.2231	0.5108	0.9163	1.6094
	x								
$C \sim U[0,4]$	0.2231	0.88844	1.7884	2.6257	3.2939	1.4195	2.4791	3.1162	3.4495
	0.5108	1.87235	3.5374	4.9421	5.9103	2.5675	4.2337	5.2879	5.6652
	0.9163	2.82716	5.1281	6.7562	7.9288	3.4847	5.6546	6.8609	7.2752
	1.6094	4.11937	6.8263	8.5792	9.8244	4.3865	7.0990	8.6043	9.6083
	$n = 50$								
	0.2231	0.46707	0.89463	1.2908	1.6311	0.70828	1.1836	1.4852	1.6391
	0.5108	0.96535	1.77927	2.4443	2.9482	1.25367	2.0572	2.5533	2.7682
	0.9163	1.43605	2.55676	3.3313	3.8224	1.70156	2.6825	3.2865	3.5232
	1.6094	2.07312	3.38458	4.2008	4.5999	2.17216	3.3065	4.0466	4.4689
	$n = 100$								
	0.2231	0.24073	0.45678	0.67363	0.82386	0.70828	1.1836	1.4852	1.6391
	0.5108	0.46615	0.87229	1.22072	1.47199	1.25367	2.0572	2.5533	2.7682
	0.9163	0.72279	1.28001	1.69584	1.92327	1.70156	2.6825	3.2865	3.5232
	1.6094	1.01154	1.69896	2.07943	2.26028	2.17216	3.3065	4.0466	4.4689
	$n = 200$								
	0.2231	0.11882	0.23321	0.32905	0.41819	0.17692	0.30286	0.38004	0.41847
	0.5108	0.23412	0.43163	0.59309	0.72899	0.31567	0.52869	0.65091	0.70707
	0.9163	0.36677	0.63919	0.82695	0.95451	0.42570	0.70554	0.84129	0.90769
	1.6094	0.50917	0.82842	1.01543	1.10098	0.54013	0.86511	1.04140	1.12257
	$n = 400$								
$C \sim U[0,3]$	0.2231	0.94896	1.9478	2.8766	3.8743	1.4393	2.5321	3.2124	3.5749
	0.5108	2.00767	3.9632	5.4971	7.3228	2.6461	4.5640	5.7794	6.6904
	0.9163	3.14946	5.7231	8.0144	11.0769	3.7311	6.3091	8.0774	10.3223
	1.6094	5.07288	8.6785	12.4421	15.1864	5.2921	8.5802	12.5194	15.9284
	$n = 50$								
	0.2231	0.50229	0.98928	1.4392	1.9000	0.73527	1.2364	1.5698	1.7607
	0.5108	0.98375	1.87897	2.6586	3.5479	1.33375	2.2041	2.7190	3.2148
	0.9163	1.56951	2.77192	3.8906	5.3132	1.86767	3.0120	3.8556	5.0151
	1.6094	2.43473	4.08571	5.7933	8.1521	2.63798	4.2494	6.0918	8.7017
	$n = 100$								
	0.2231	0.26292	0.50400	0.7239	0.93641	0.37796	0.63105	0.78854	0.87771
	0.5108	0.51303	0.95829	1.3515	1.77883	0.66910	1.11701	1.38841	1.57500
	0.9163	0.78172	1.44738	1.9753	2.62668	0.92451	1.53323	1.93982	2.37176
	1.6094	1.23053	2.09186	2.9207	4.66967	1.26364	2.09036	2.93411	5.18373
	$n = 200$								
	0.2231	0.12854	0.24391	0.35276	0.46193	0.18359	0.31086	0.38834	0.43965
	0.5108	0.25551	0.47982	0.66380	0.84978	0.32215	0.55223	0.69143	0.79304
	0.9163	0.39291	0.71491	0.97562	1.29319	0.45197	0.75955	0.96646	1.18841
	1.6094	0.60111	1.04720	1.39592	2.68882	0.62569	1.03798	1.43771	2.98205
	$n = 400$								

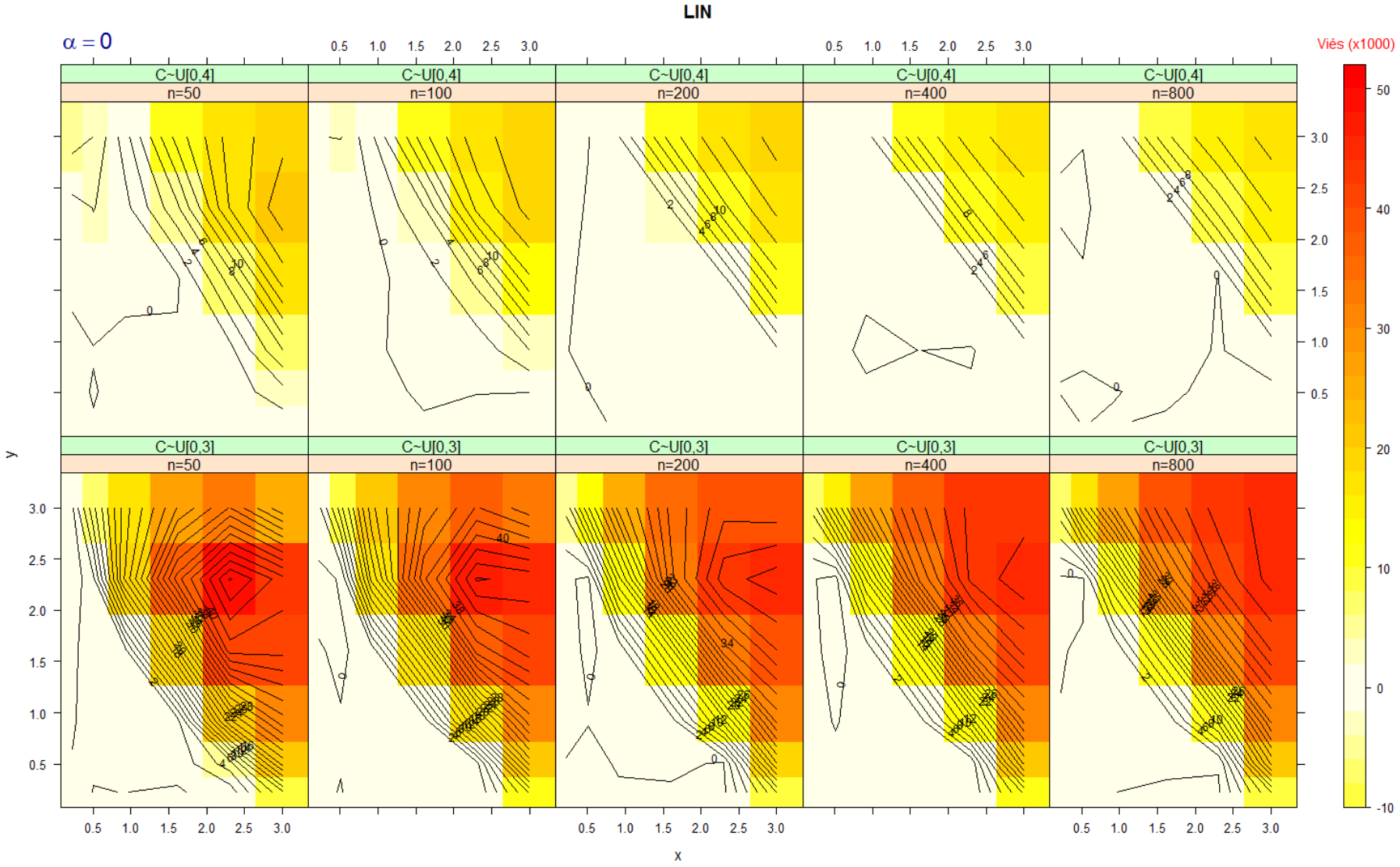


Figura D.1: Enviesamento de $\widehat{F}_{12}(x, y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 0$).

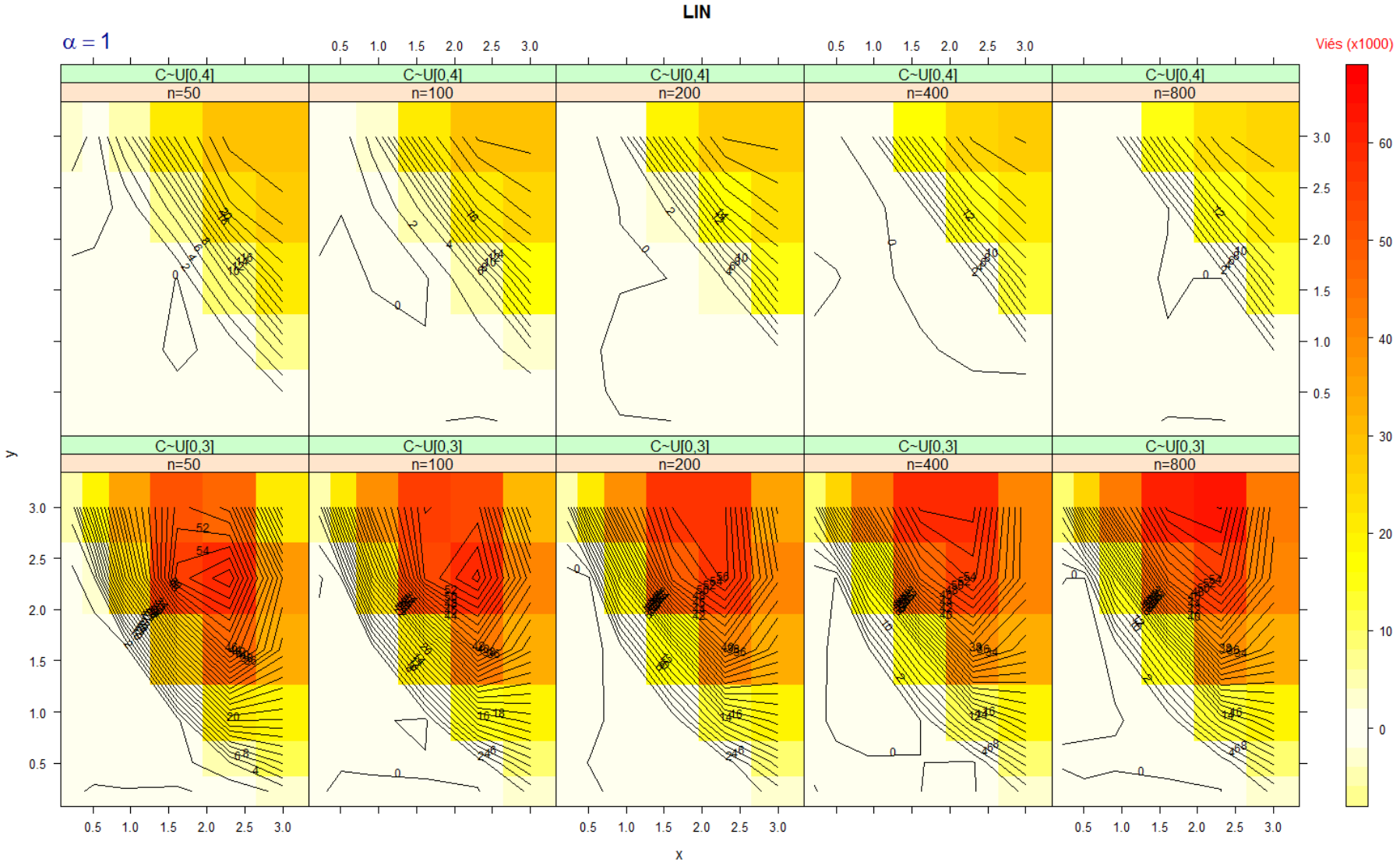


Figura D.2: Enviesamento de $\widehat{F}_{12}(x, y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 1$).

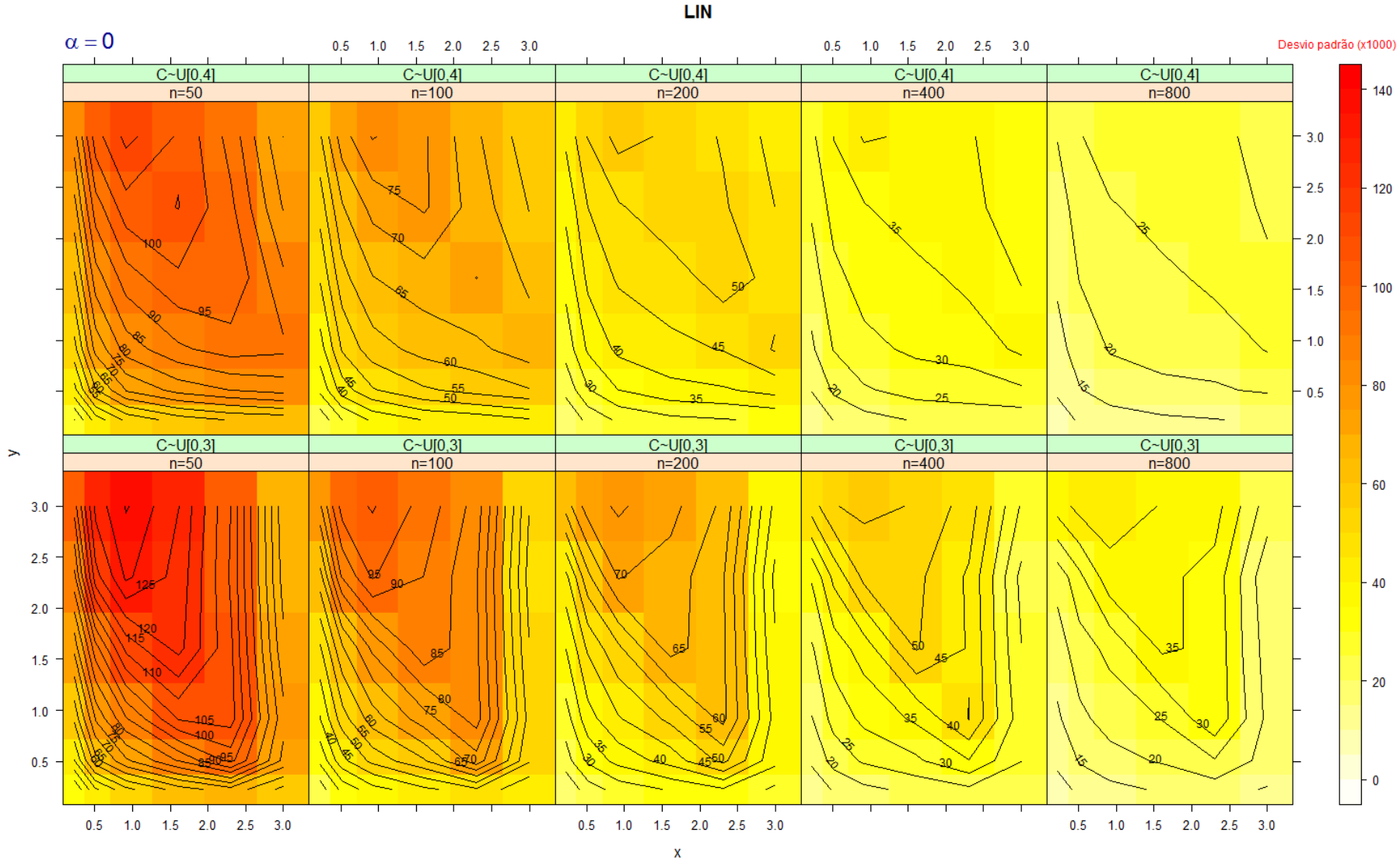


Figura D.3: Desvio padrão de $\widehat{F}_{12}(x, y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 0$).

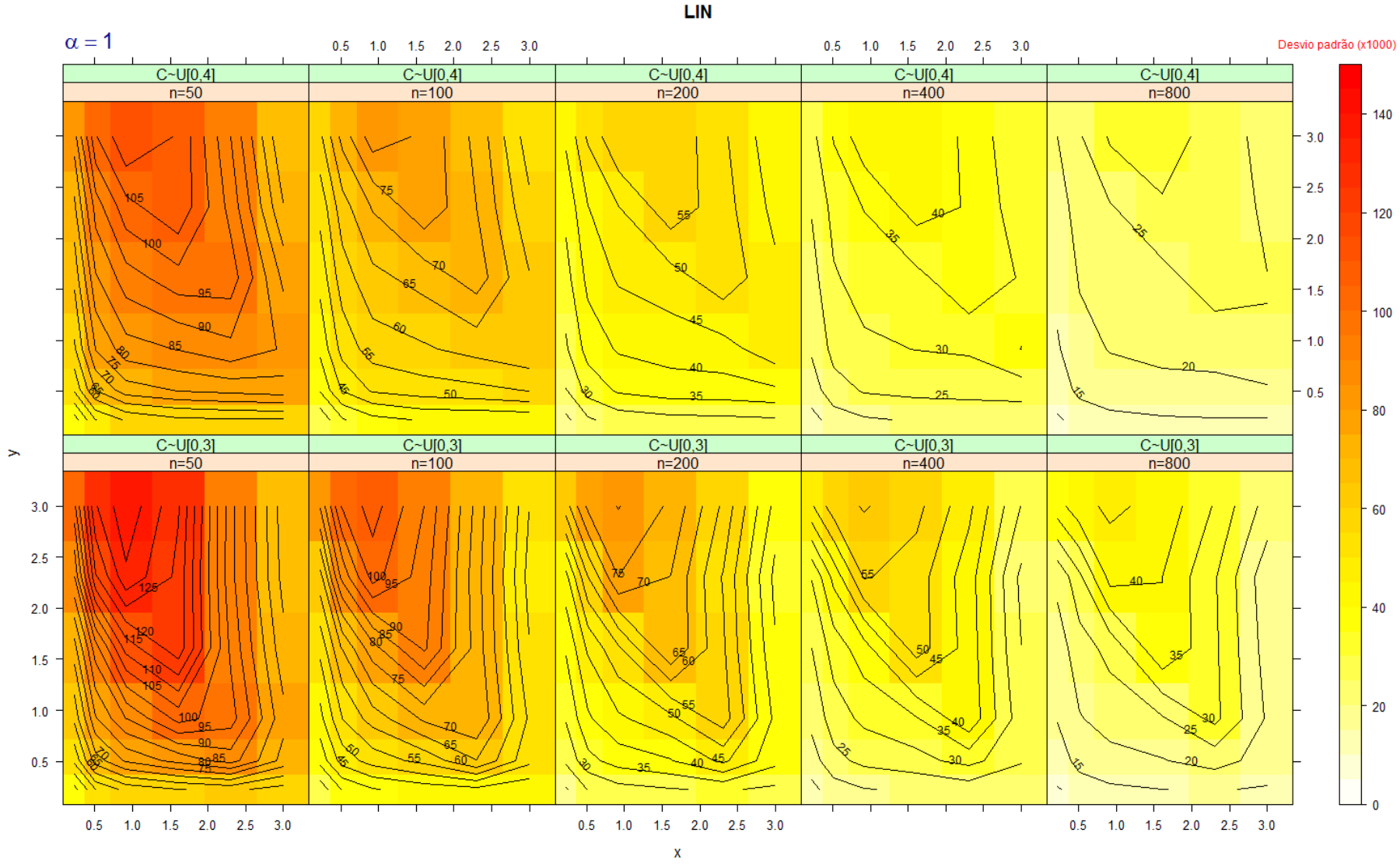


Figura D.4: Desvio padrão de $\widehat{F}_{12}(x, y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 1$).

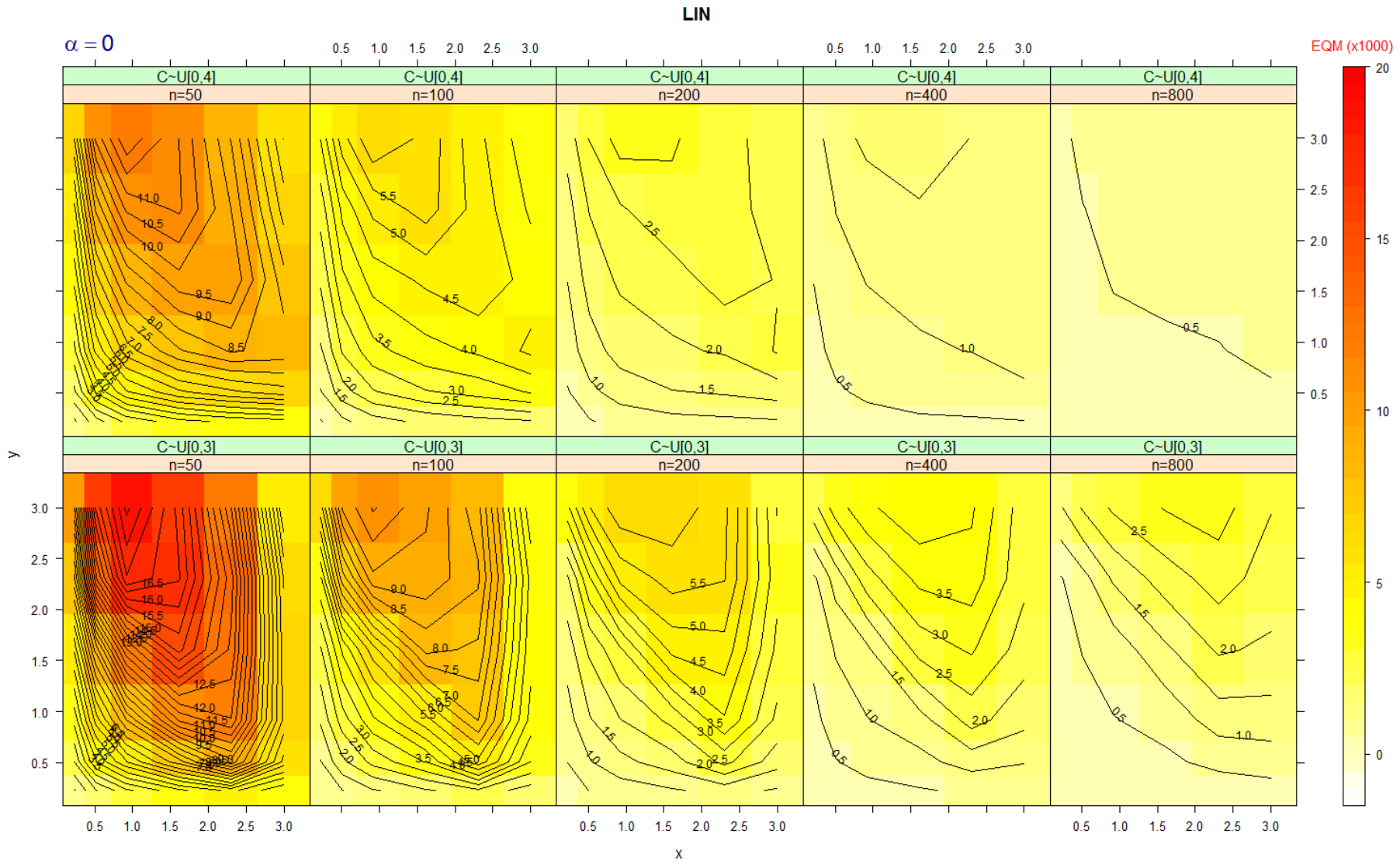


Figura D.5: Erro quadrático médio de $\widehat{F}_{12}(x, y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 0$).

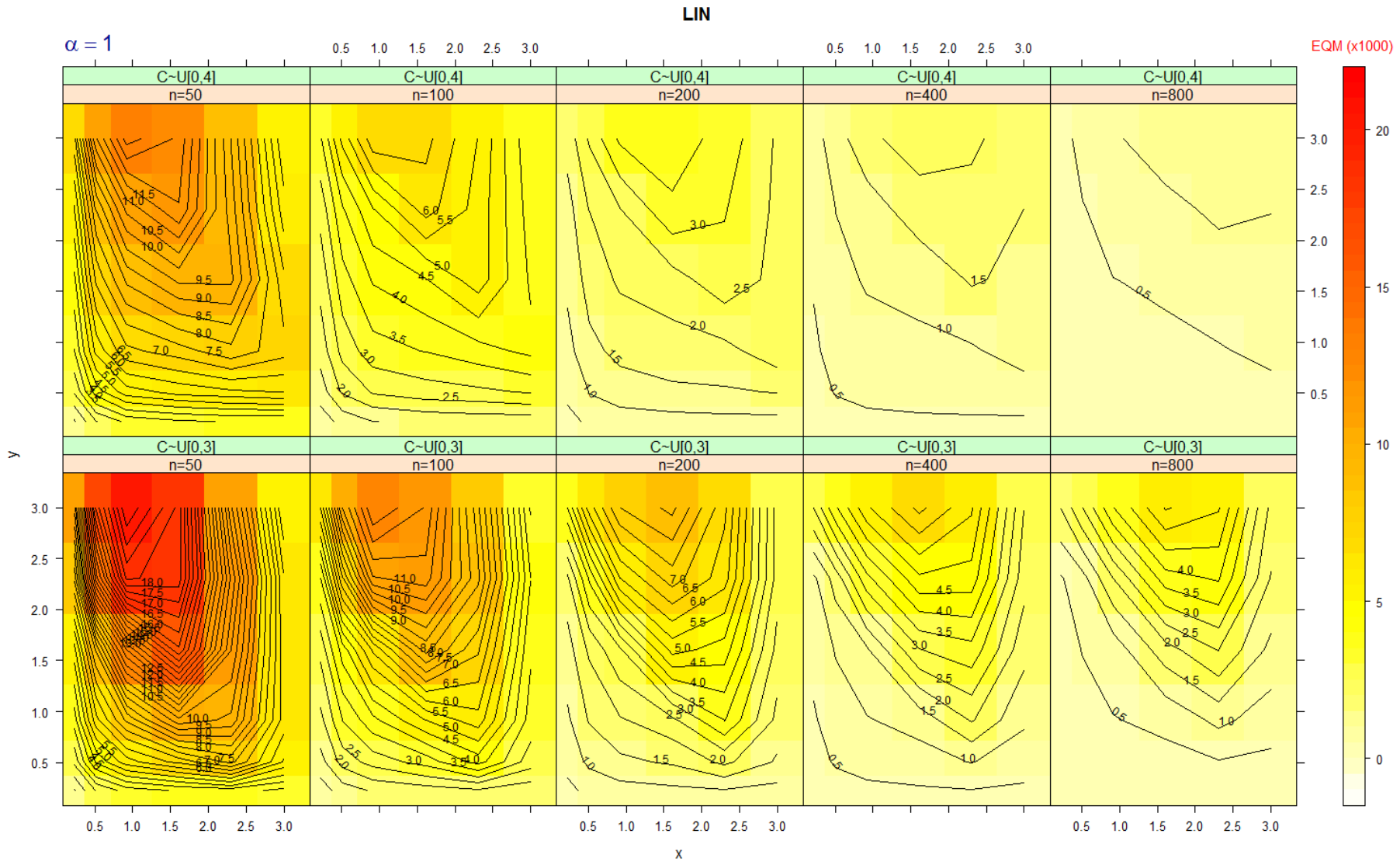


Figura D.6: Erro quadrático médio de $\widehat{F}_{12}(x, y)$ (x1000) obtido da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 1$).

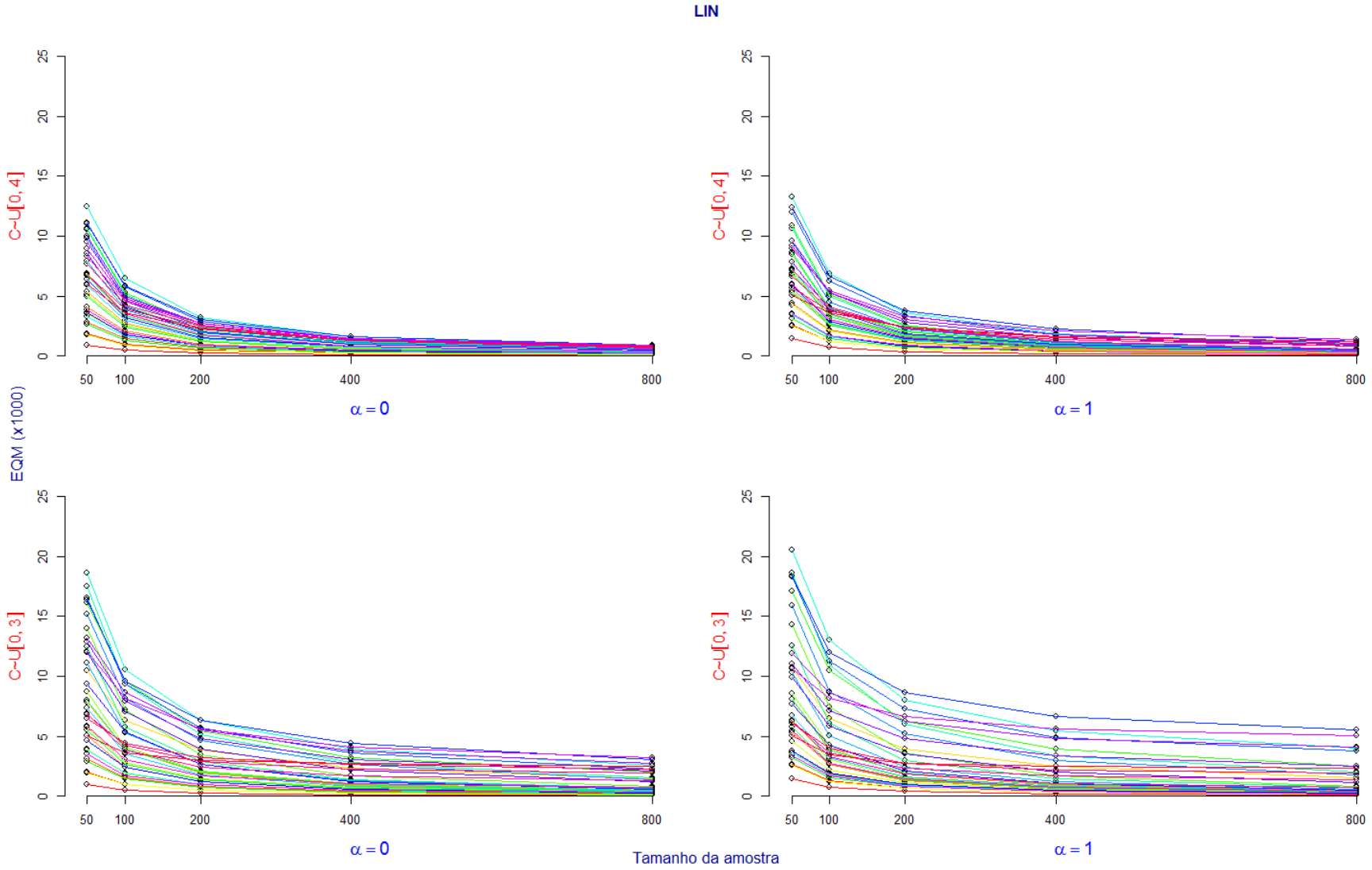


Figura D.7: Erro quadrático médio de $\widehat{F}_{12}(x, y)$ (x1000) versus o tamanho da amostra.

Anexo E Eficiência relativa

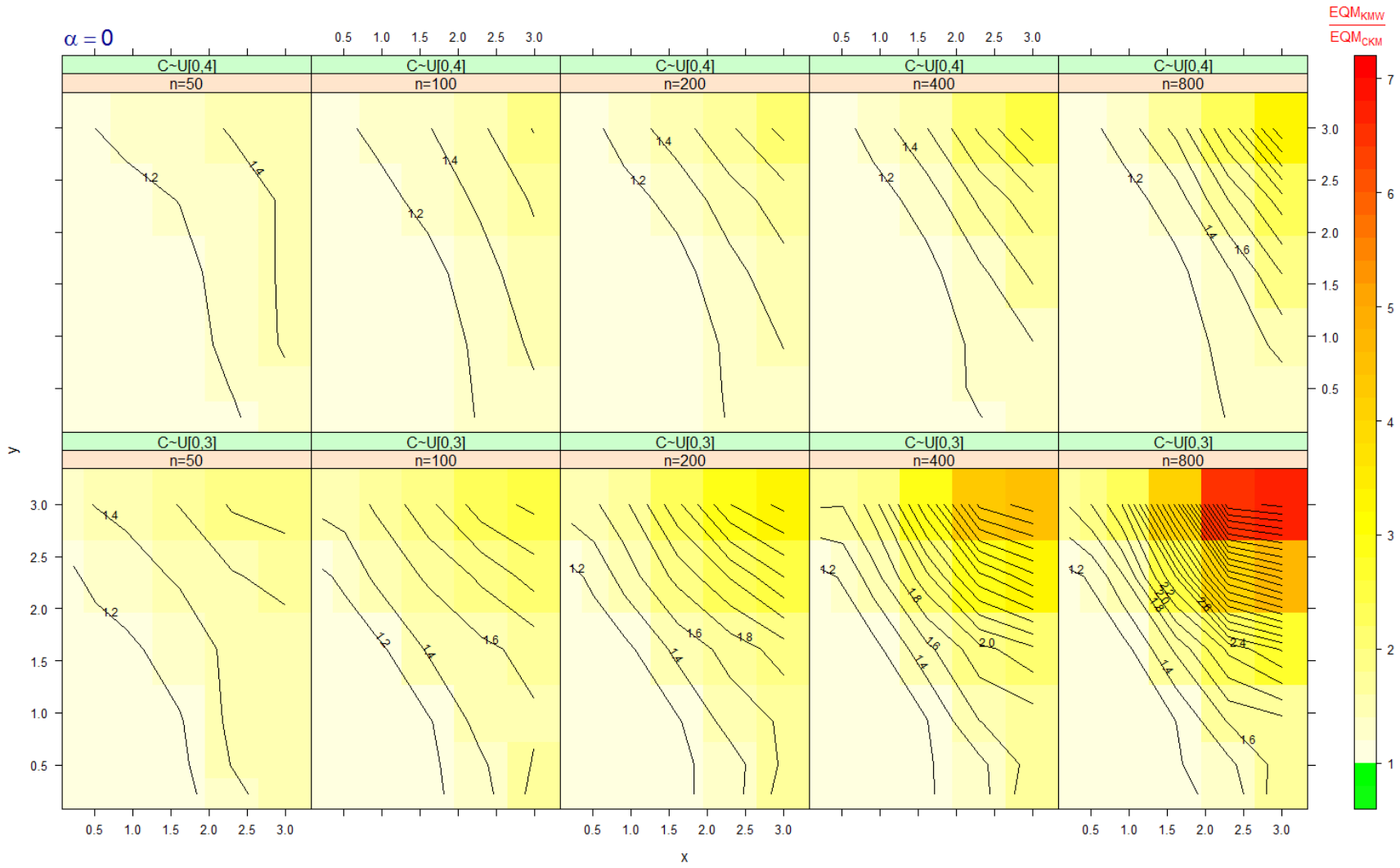


Figura E.1: Eficiência do estimador KMW relativa ao estimador CKM, obtida da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 0$).

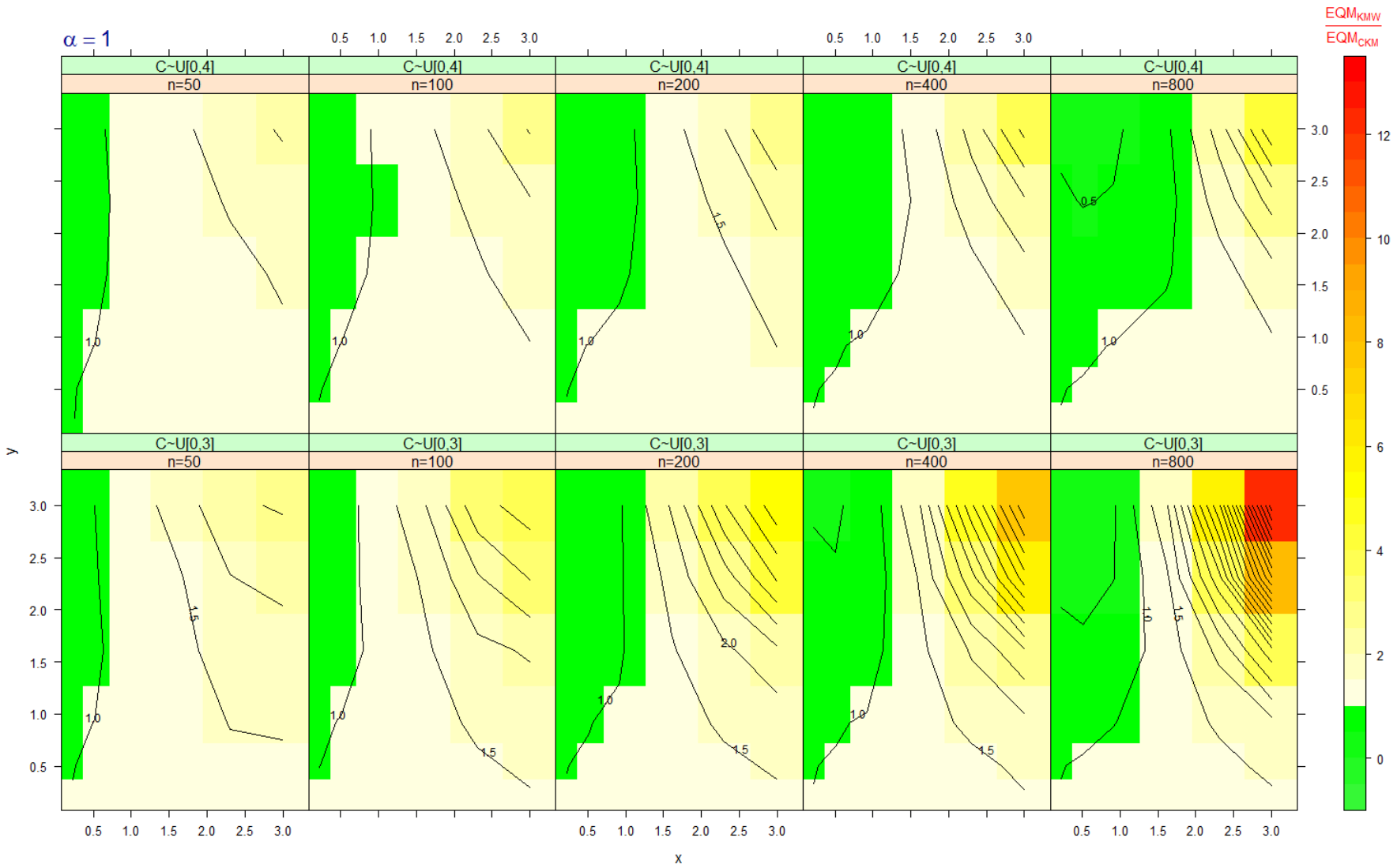


Figura E.2: Eficiência do estimador KMW relativa ao estimador CKM, obtida da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 1$).

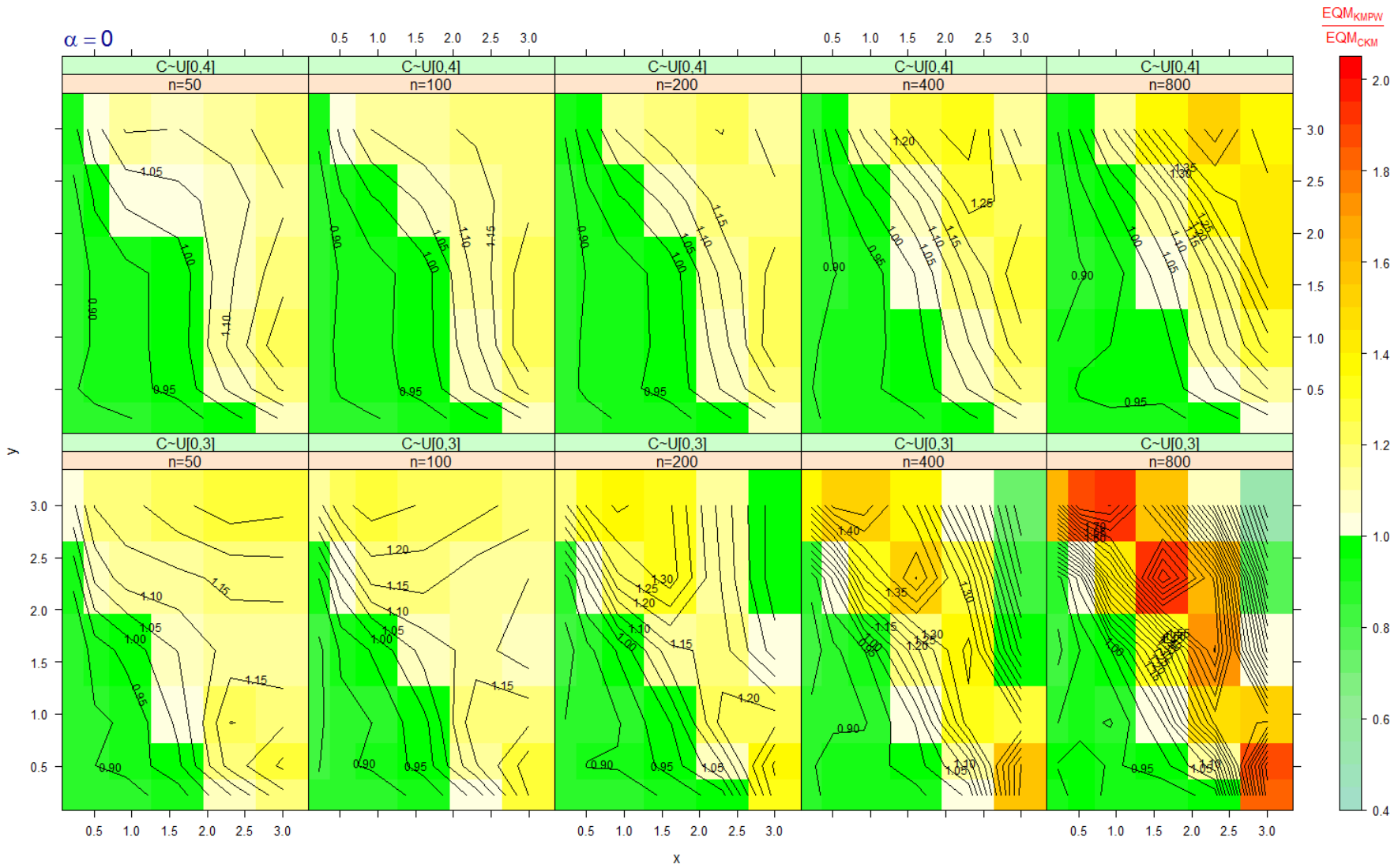


Figura E.3: Eficiência do estimador KMPW relativa ao estimador CKM, obtida da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 0$).

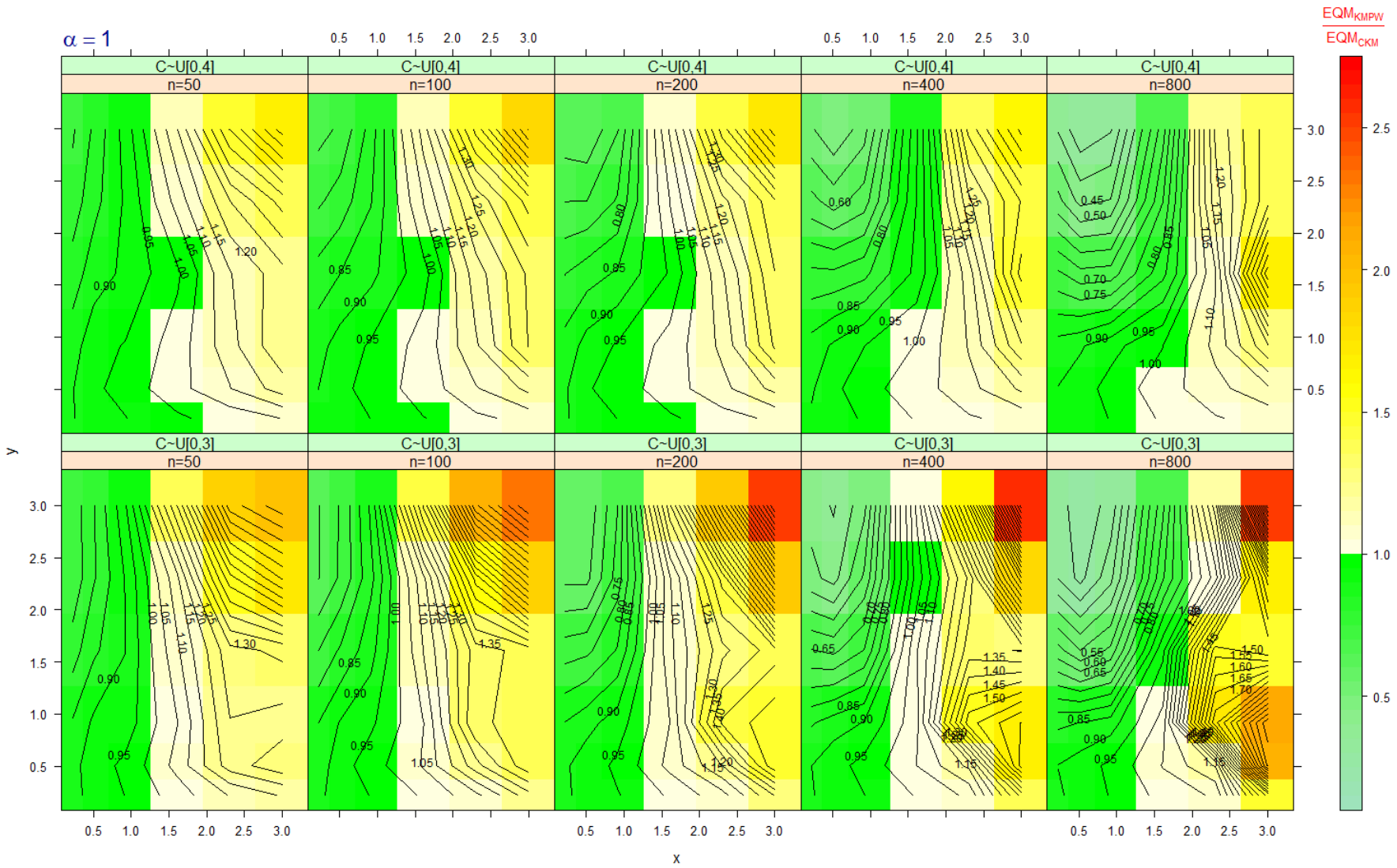


Figura E.4: Eficiência do estimador KMPW relativa ao estimador CKM, obtida da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 1$).

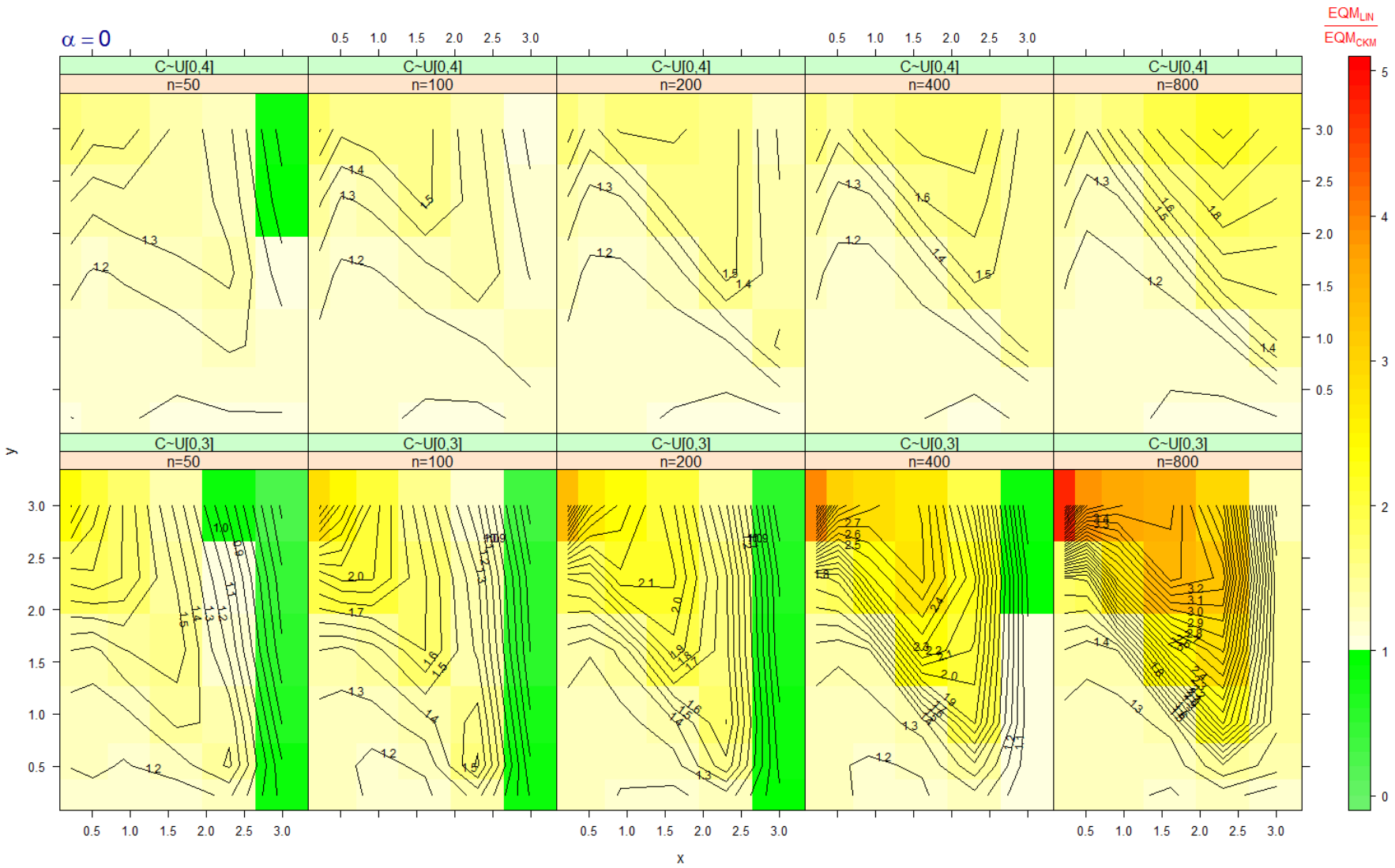


Figura E.5: Eficiência do estimador LIN relativa ao estimador CKM, obtida da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 0$).

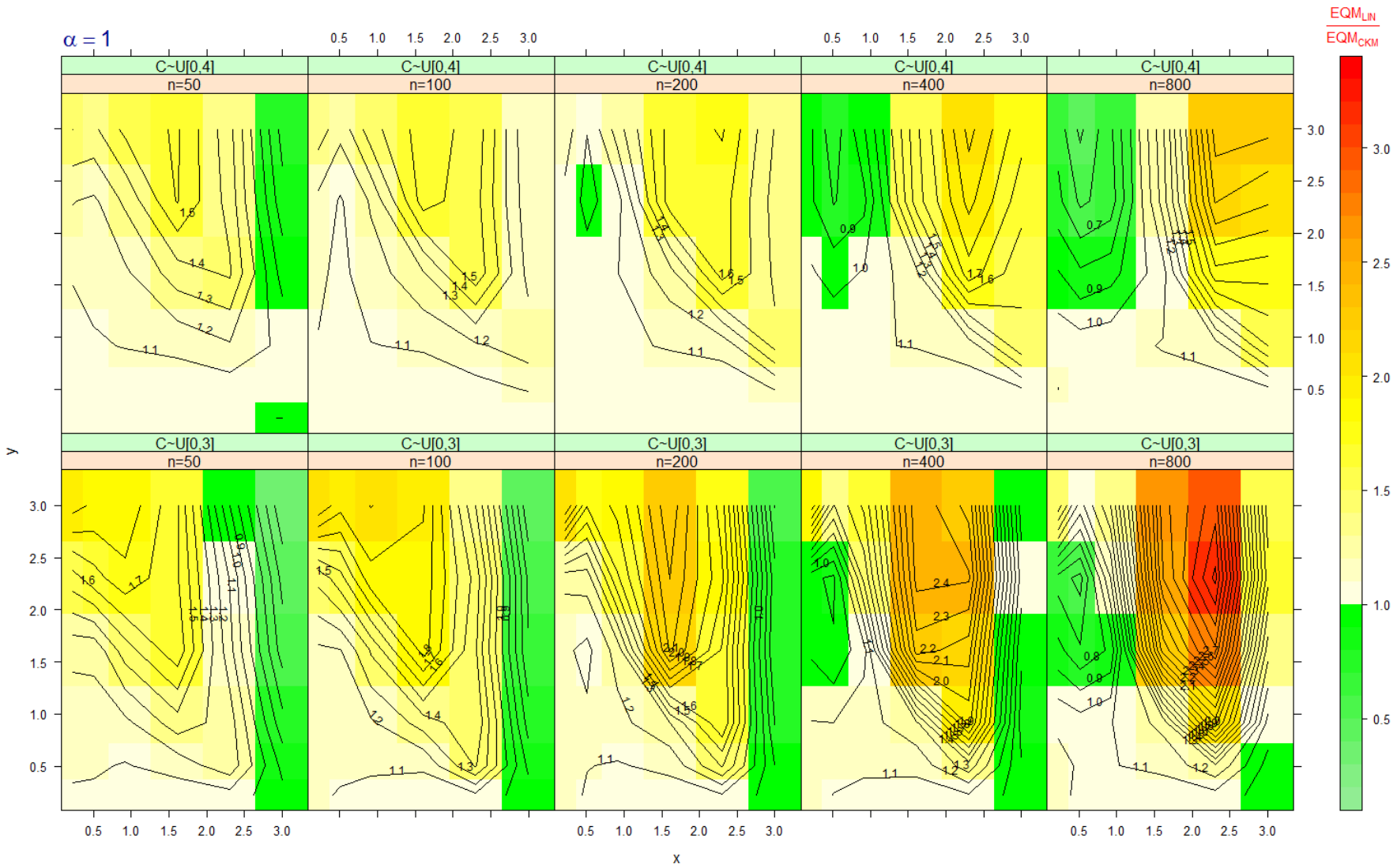


Figura E.6: Eficiência do estimador LIN relativa ao estimador CKM, obtida da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 1$).

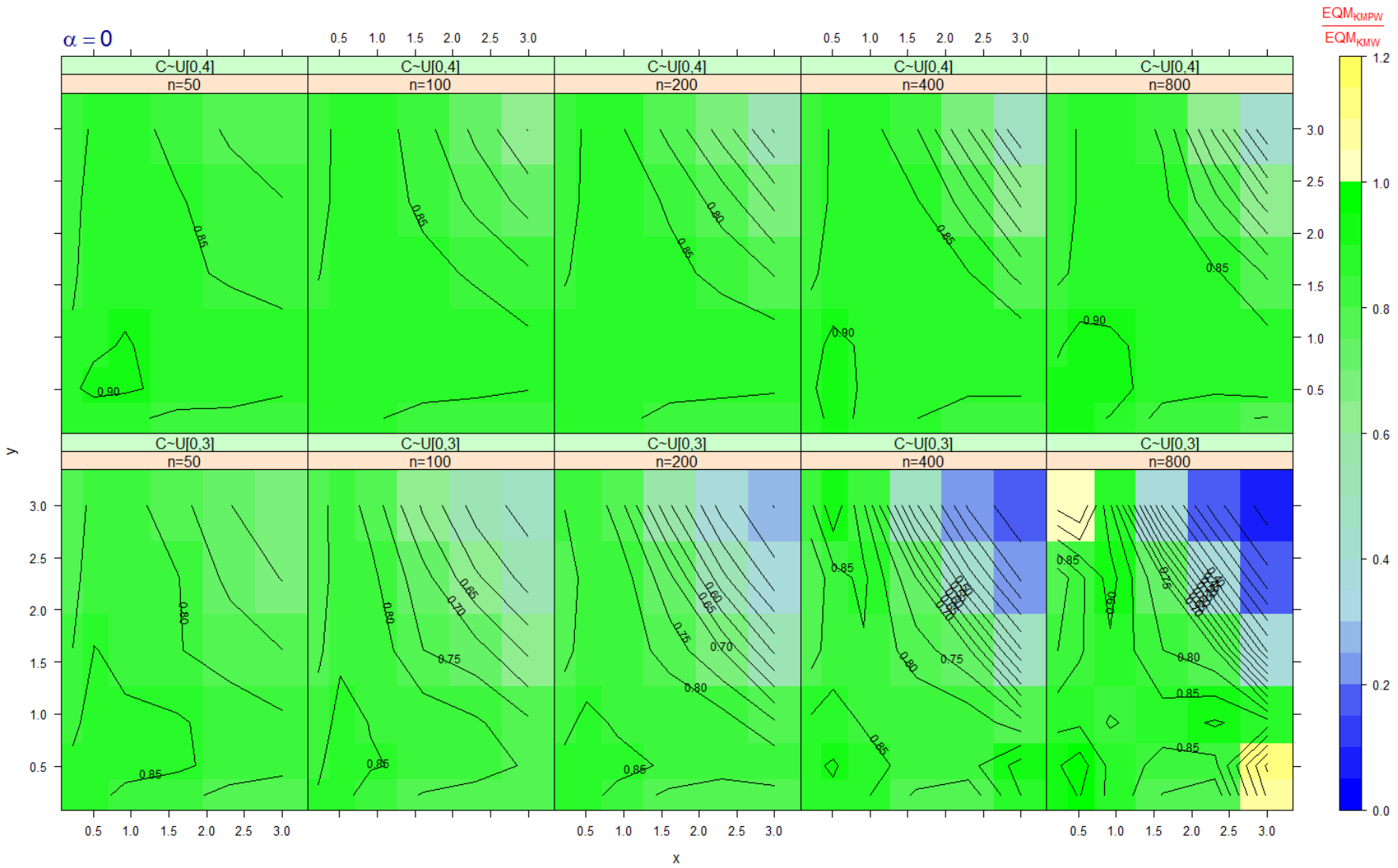


Figura E.7: Eficiência do estimador KMPW relativa ao estimador KMW, obtida da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 0$).

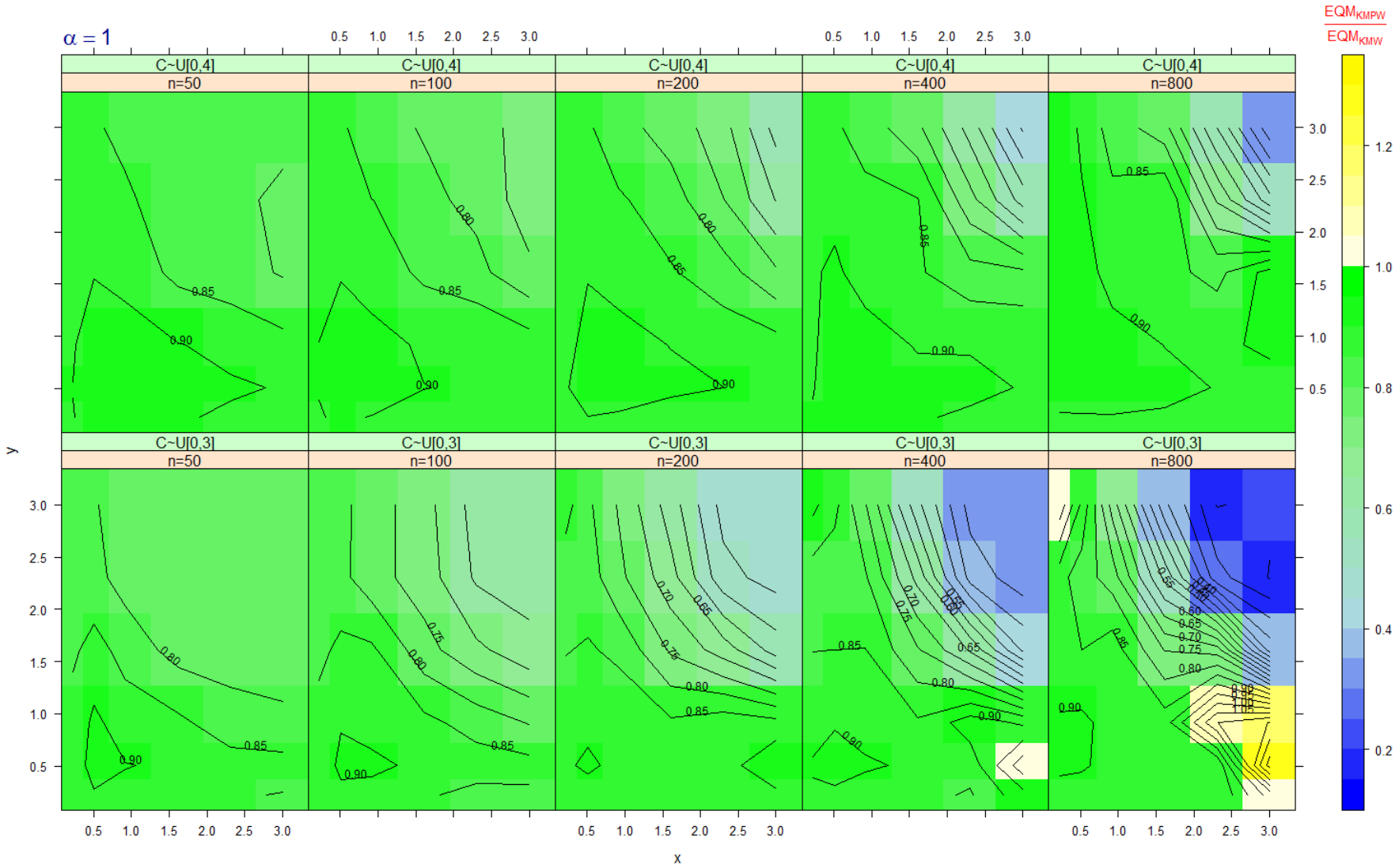


Figura E.8: Eficiência do estimador KMPW relativa ao estimador KMW, obtida da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 1$).

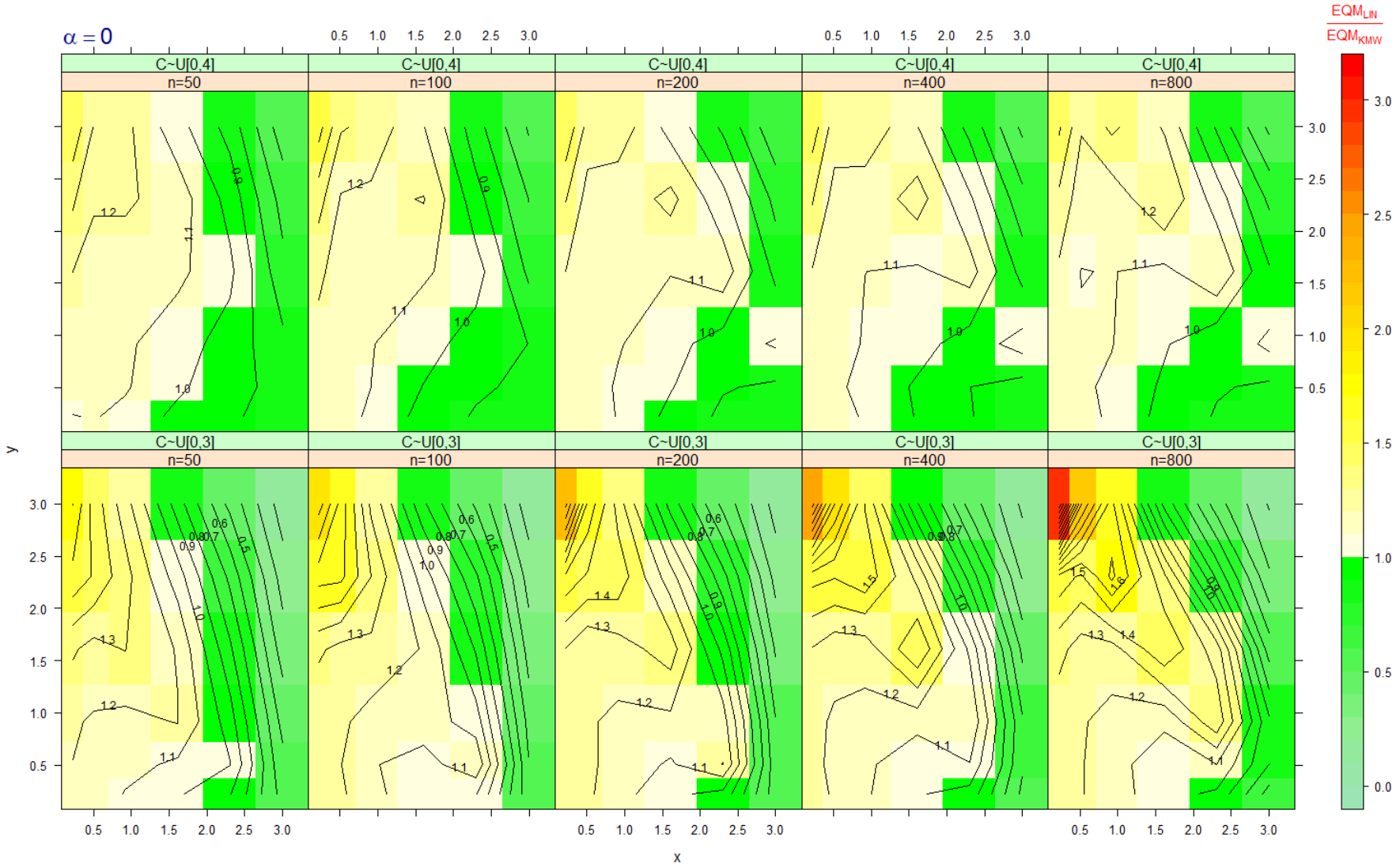


Figura E.9: Eficiência do estimador LIN relativa ao estimador KMW, obtida da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 0$).

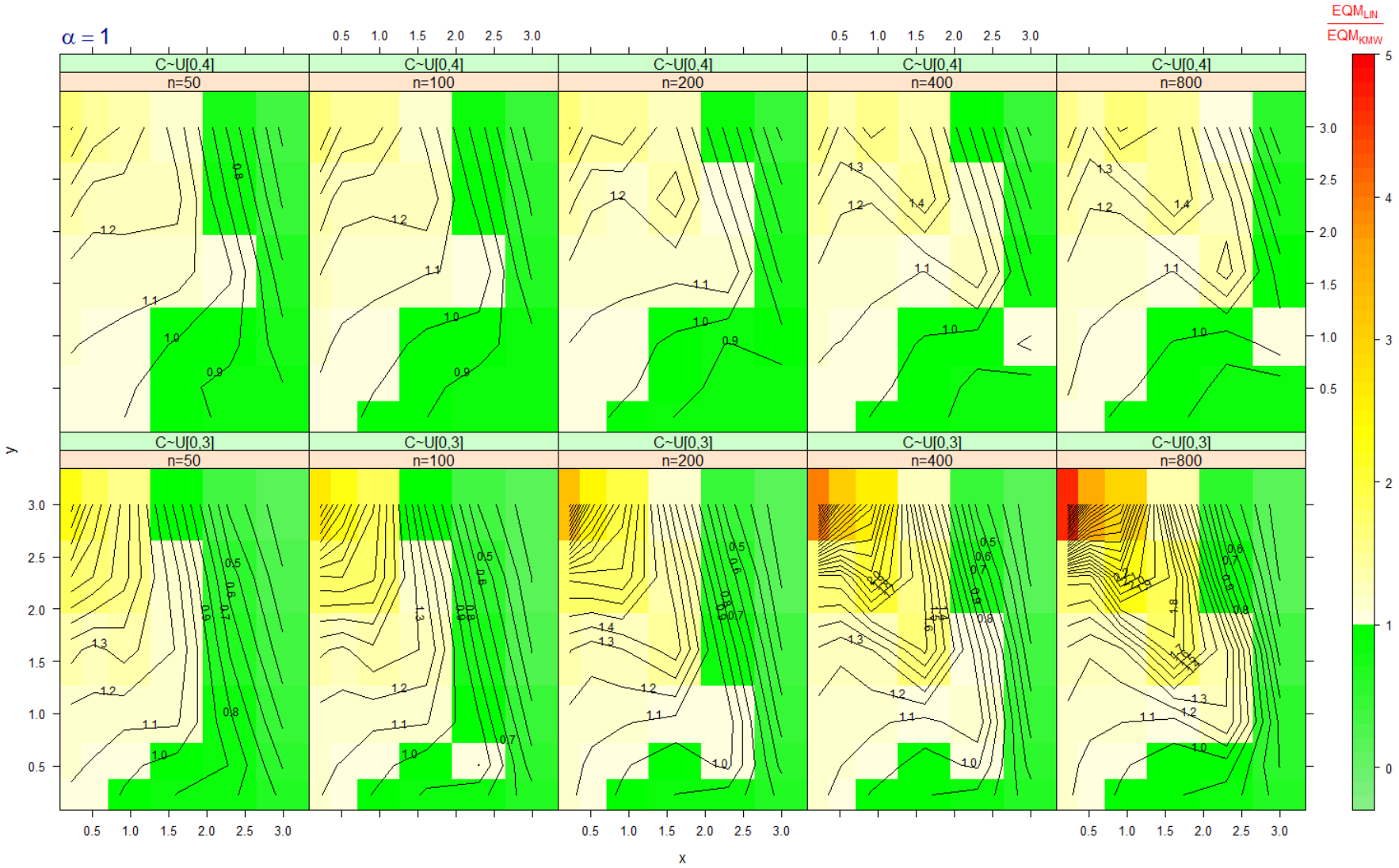


Figura E.10: Eficiência do estimador LIN relativa ao estimador KMW, obtida da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 1$).

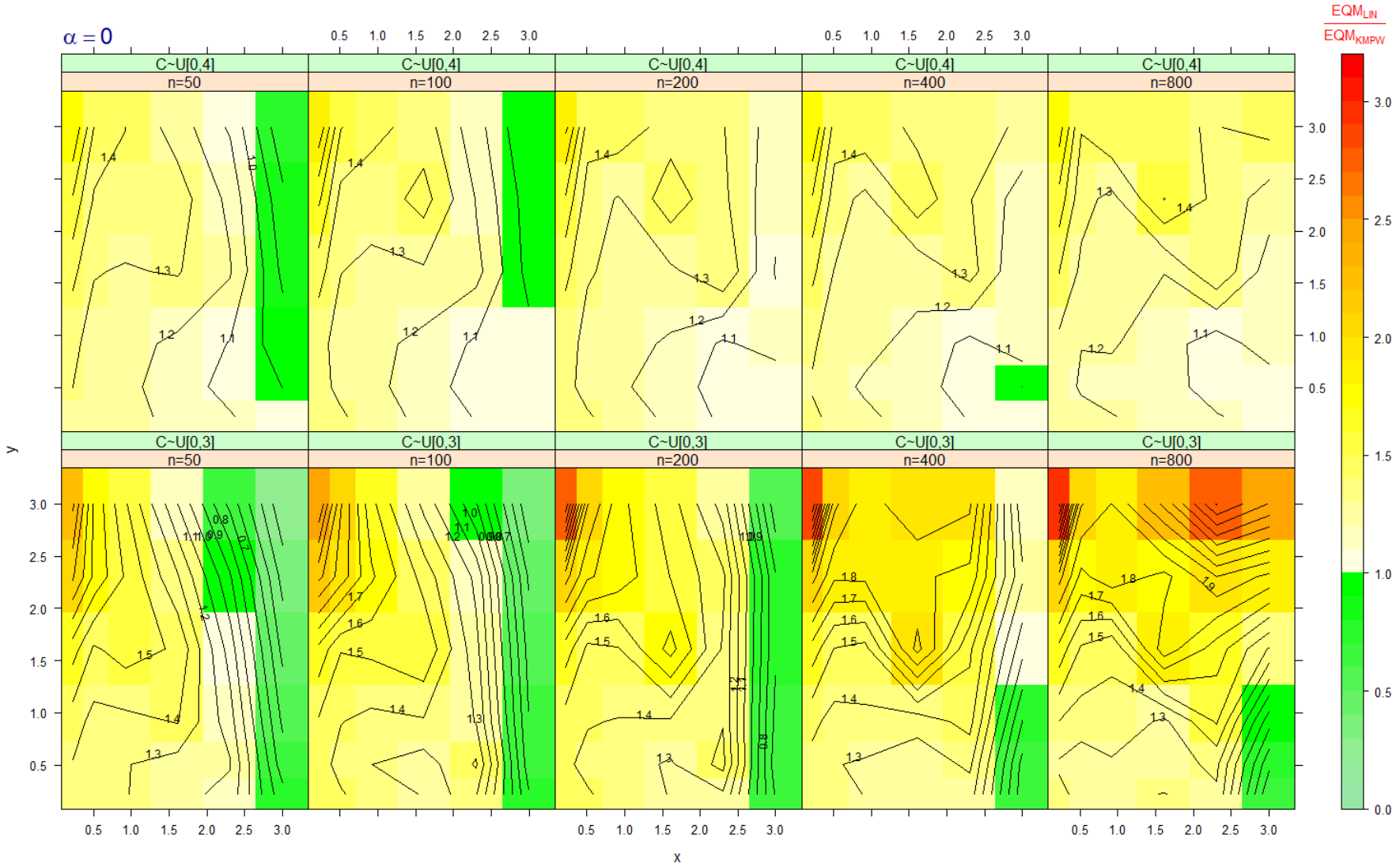


Figura E.11: Eficiência do estimador LIN relativa ao estimador KMPW, obtida da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 0$).

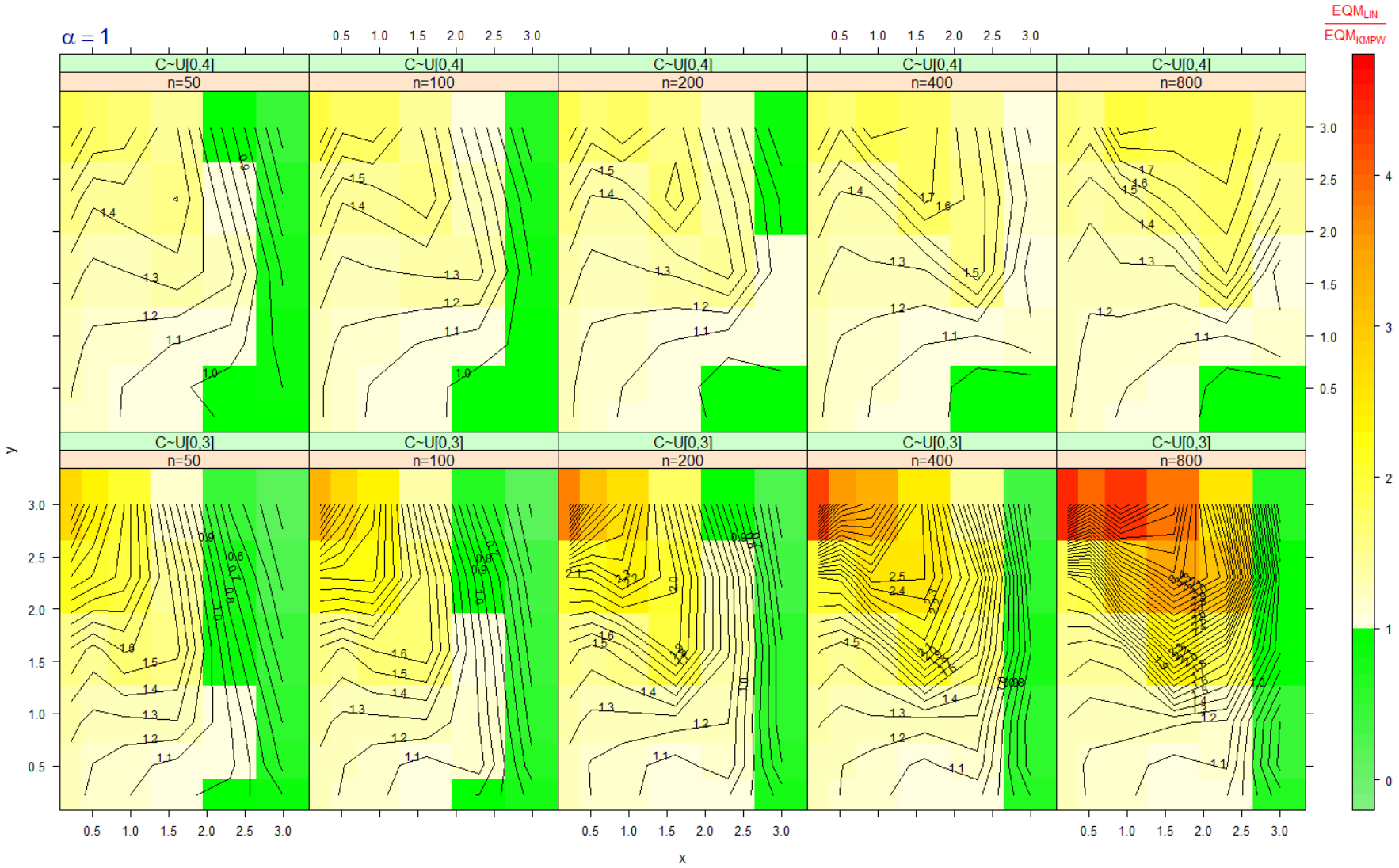


Figura E.12: Eficiência do estimador LIN relativa ao estimador KMPW, obtida da simulação de 10000 amostras quando x e y são sujeitos a censura aleatória pela direita ($\alpha = 1$).

Bibliografia

Borgan, O. (2005). Kaplan-Meier Estimator. Em P. Armitage, & T. Colton (Edits.), *Encyclopedia of Biostatistics* (2nd ed.). Hoboken, New Jersey, United States of America: John Wiley & Sons.

Braun, W. J., & Murdoch, D. J. (2007). *A First Course in Statistical Programming with R.* Cambridge, England, United Kingdom: Cambridge University Press.

Chapman, B., Jost, G., & van der Pas, R. (2008). *Using OpenMP: Portable Shared Memory Parallel Programming.* Cambridge, Massachusetts, United States of America: The MIT Press.

Crawley, M. J. (2007). *The R Book.* Chichester, England, United Kingdom: John Wiley & Sons.

de Uña-Álvarez, J., & Amorim, A. P. (Fevereiro de 2011). A semiparametric estimator of the bivariate distribution function for censored gap times. *Biometrical Journal, 53*(1), 113-127. doi:10.1002/bimj.201000063

de Uña-Álvarez, J., & Meira-Machado, L. F. (15 de Outubro de 2008). A simple estimator of the bivariate distribution function for censored gap times. *Statistics and Probability Letters, 78*(15), 2440-2445. doi:10.1016/j.spl.2008.02.031

Dikta, G. (1 de Março de 1998). On Semiparametric random censorship models. *Journal of Statistical Planning and Inference, 66*(2), 253-279. doi:10.1016/S0378-3758(97)00091-8

Dobson, A. J., & Barnett, A. G. (2008). *An Introduction to Generalized Linear Models* (3rd ed.). Boca Raton, Florida, United States of America: Chapman & Hall/CRC.

Efron, B. (1982). *The Jacknife, the Bootstrap and Other Resampling Plans* (Vol. 38). Philadelphia, Pennsylvania, United States of America: Society for Industrial and Applied Mathematics.

Efron, B., & Tibshirani, R. J. (1994). *An Introduction to the Bootstrap.* Boca Raton, Florida, United States of America: Chapman & Hall/CRC.

Hardle, W., & Simar, L. (2007). *Applied Multivariate Statistical Analysis* (2nd ed.). Berlin, Germany: Springer-Verlag.

Hosmer, Jr., D. W., & Lemeshow, S. (1999). *Applied Survival Analysis: Regression Modelling of Time to Event Data.* New York, New York, United States of America: John Wiley & Sons.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data.* New York, New York, United States of America: Springer-Verlag.

Johnson, M. E. (1987). *Multivariate Statistical Simulation.* New York, New York, United States of America: John Wiley & Sons.

Johnson, R. A., Evans, J. W., & Green, D. W. (1999). *Some bivariate distributions for modelling the strength properties of lumber.* Madison, Wisconsin: United States Departement of Agriculture, Forest Service, Forest Products Laboratory. Obtido de http://www2.fpl.fs.fed.us/documnts/fplrp/fplrp575.pdf

Jones, B. L., & Aitken, P. (2003). *Sams Teach Yourself C in 21 Days* (6th ed.). Indianapolis, Indiana, United States of America: Sams Publishing.

Kaplan, E. L., & Meier, P. (Junho de 1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association, 53*(282), 457-481. doi:10.1080/01621459.1958.10501452

Klein, J. P., & Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data* (2nd ed.). New York, New York, United States of America: Springer-Verlag.

Knight, K. (2000). *Mathematical Statistics.* Boca Raton, Florida, United States of America: Chapman & Hall/CRC.

Kochan, S. G. (2005). *Programming in C* (3rd ed.). Indianapolis, Indiana, United States of America: Sams Publishing.

Kotz, S., Balakrishnan, N., & Johnson, N. L. (2000). *Continuous Multivariate Distributions* (2nd ed., Vol. 1). New York, New York, United States of America: John Wiley & Sons.

Lee, E. T., & Wang, W. J. (2003). *Statistical Methods for Survival Data Analysis* (3rd ed.). Hoboken, New Jersey, United States of America: John Wiley & Sons.

Lehmann, E. L., & Casella, G. (1998). *Theory of Point Estimation* (2nd ed.). New York, New York, United States of America: Springer-Verlag.

Lin, D. Y., Sun, W., & Ying, Z. (1999). Nonparametric estimation of the gap time distributions for serial events with censored data. *Biometrika, 86*(1), 59-70. doi:10.1093/biomet/86.1.59

Lu, J.-C., & Bhattacharyya, G. K. (1990). Some New Constructions of Bivariate Weibull Models. *Annals of the Institute of Statistical Mathematics, 42*(3), 543-559. doi:10.1007/BF00049307

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models* (2nd ed.). Boca Raton, Florida, United States of America: Chapman & Hall/CRC.

Mittal, H. V. (2011). *R Graphs Cookbook.* Birmingham, England, United Kingdom: Packt Publishing.

Montgomery, D. C., & Runger, G. C. (2011). *Applied Statistics and Probability for Engineers* (5th ed.). Hoboken, New Jersey, United States of America: John Wiley & Sons.

Moreira, A., & Meira-Machado, L. (7 de Março de 2012). survivalBIV: Estimation of the Bivariate Distribution Function for Sequentially Ordered Events Under Univariate Censoring. *Journal of Statistical Software, 46*(13), 1-16. Obtido de http://www.jstatsoft.org/v46/i13/

Moreira, A., Araújo, A. A., & Machado, L. M. (2012). survivalBIV: Estimation of the bivariate distribution function. (R package version 1.4). Obtido de http://CRAN.R-project.org/package=survivalBIV

Murrell, P. (2006). *R Graphics.* Boca Raton, Florida, United States of America: Chapman & Hall/CRC.

OpenMP Architecture Review Board. (Maio de 2008). OpenMP Application Program Interface. (Version 3.0). OpenMP Architecture Review Board. Obtido de http://www.openmp.org/mp-documents/spec30.pdf

OpenMP Architecture Review Board. (Novembro de 2008). Summary of OpenMP 3.0 C/C++ Syntax. OpenMP Architecture Review Board. Obtido de http://www.openmp.org/mp-documents/OpenMP3.0-SummarySpec.pdf

Pestana, D. D., & Velosa, S. F. (2008). *Introdução à Probabilidade e à Estatística* (3ª ed.). Lisboa, Portugal: Fundação Calouste Gulbenkian.

Press, S. J. (2005). *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference* (2nd ed.). Mineola, New York, United States of America: Dover Publications.

R Core Team. (2012). R Installation and Administration. Vienna, Austria: R Foundation for Statistical Computing. Obtido de http://cran.r-project.org/doc/manuals/R-admin.pdf

R Core Team. (2012). R Internals. Vienna, Austria: R Foundation for Statistical Computing. Obtido de http://cran.r-project.org/doc/manuals/R-ints.pdf

R Core Team. (2012). R Language Definition. Vienna, Austria: R Foundation for Statistical Computing. Obtido de http://cran.r-project.org/doc/manuals/R-lang.pdf

R Core Team. (2012). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Obtido de http://www.R-project.org/

R Core Team. (2012). Writing R Extensions. Vienna, Austria: R Foundation for Statistical Computing. Obtido de http://cran.r-project.org/doc/manuals/R-exts.pdf

Reis, E. (2001). *Estatística Multivariada Aplicada* (2ª ed.). Lisboa, Portugal: Edições Sílabo.

Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R.* New York, New York, United States of America: Springer-Verlag. doi:10.1007/978-0-387-75969-2

Satten, G. A., & Datta, S. (Agosto de 2001). The Kaplan-Meier Estimator as an Inverse-Probability-of-Censoring Weighted Average. *The American Statistician, 55*(3), 207-210. doi:10.1198/000313001317098185

Venables, W. N., Smith, D. M., & R Core Team. (2012). An Introduction to R. Vienna, Austria: R Foundation for Statistical Computing. Obtido de http://cran.r-project.org/doc/manuals/R-intro.pdf

Wasserman, L. (2006). *All of Nonparametric Statistics.* New York, New York, United States of America: Springer-Verlag.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R.* Boca Raton, Florida, United States of America: Chapman & Hall/CRC.